IOWA STATE UNIVERSITY
Digital Repository

Retrospective Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2000

Efficiency of Markov chain Monte Carlo algorithms for Bayesian inference in random regression models

Ho Huei Liu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

 Part of the Animal Sciences Commons, Biostatistics Commons, Physiology Commons, and the Veterinary Physiology Commons

Recommended Citation
Liu, Ho Huei, "Efficiency of Markov chain Monte Carlo algorithms for Bayesian inference in random regression models " (2000). *Retrospective Theses and Dissertations*. 12347.
https://lib.dr.iastate.edu/rtd/12347

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Efficiency of Markov chain Monte Carlo algorithms for Bayesian inference in random regression models

by

Ho Huei Liu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Majors: Statistics; Animal Breeding and Genetics

Major Professors: Hal S. Stern and Jack C. M. Dekkers

Iowa State University

Ames, Iowa

2000

Copyright © Ho Huei Liu, 2000. All rights reserved.

UMI Number: 9990472

UMI®

UMI Microform9990472
Copyright 2001 by Bell & Howell Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Ho Huei Liu

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

When a character varies over time, successive measurements are required if we are interested in studying this variation. For example, longitudinal data on animal growth traits or dairy cow lactation traits may be collected with measurements taken every few days. Figure 1.1 displays an example, pig weight over time.

It is natural to expect that repeated measurements on the same animal are correlated. The correlation between measurements may decrease as the time interval between them gets larger. The correlation between two measurements on the same animal is generally higher than two measurements on different animals. Every animal has its own unique genetic characteristics which partly determine its growth profile. Animals differentially express their genetic potential in different environments or in different life stages. As a result, the measurements or phenotypic values used as a measure of performance for a trait on an animal are in general determined by genetic components, environmental factors, their interactions, and sampling errors. An animal's "breeding value" is represented by its genetic potential or genetic value, as it is this aspect that can be partially passed on to its offspring.

A common approach to selection of animals for breeding with reference to growth traits is to model the repeated measurements as functions of parameters describing genetic and environmental factors. Many mathematical models have been proposed to describe the shape of growth curves of animals. The best model depends on the nature of the data in practice. For example, different curves have been used to model milk yield data in different countries, each optimized to the relevant lactation curves (Jamrozik et

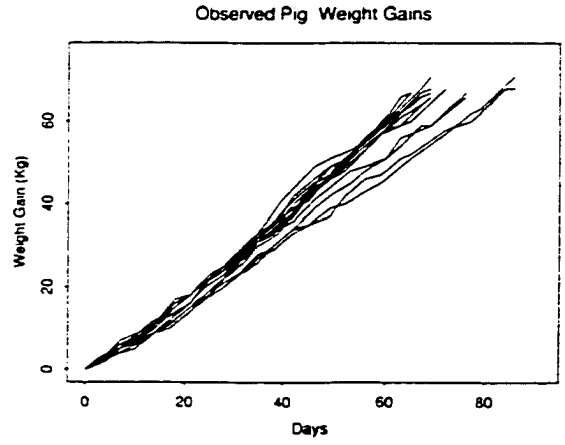

Figure 1.1 Plot showing pig weights gains for 12 different pigs over time

al., 1997).

Linear mixed models have been used in animal science since they were introduced by Henderson (1950). For longitudinal data, the error terms in the models are assumed to be either independent or correlated with a structural pattern, depending on the assumed relationship between repeated measurements. As the structure of covariation becomes more complicated, a larger number of variance components is required. This results in an increased computational burden and in an increase in the required size of data for estimating variance parameters.

Given a specific linear model, there are two major approaches to inference. REML-BLUP (restricted maximum likelihood - best linear unbiased predictor) methods are commonly used to estimate the contribution of different sources of variance and to evaluate the breeding value of individual animal. The variance of the REML-BLUP estimates can be obtained via large sample arguments (for REML estimates of variance components) and analytic arguments (for BLUPs of animal values conditional on vari-

ance components) (Patterson and Thompson, 1971; Harville, 1977). Variance estimates for BLUPs that account for uncertainty in the variance components are possible (Kackar and Harville, 1984), but not generally used.

A Bayesian approach to inference provides an alternative to the traditional REML-BLUP approach. There all parameters are treated as random variables with inferences for parameters derived from the posterior distribution of model parameters given the observed data (Gianola and Fernando, 1986). In this way uncertainty about all parameters is addressed. Two somewhat difficult issues that arise in carrying out a Bayesian analysis are the specification of prior distributions for model parameters (Berger, 1985), and techniques for evaluating (or simulating) from the posterior distribution (Gilks, 1996; Gilks and Roberts, 1996; Bennett, 1996).

Often the posterior distribution is studied via simulation using Markov chain Monte Carlo (MCMC) methods, e.g., the Gibbs sampling algorithm (Geman and Geman, 1984) or the Metropolis-Hastings algorithm (Metropolis, 1953; Hastings, 1970; Chib and Greenberg, 1995). An important issue then is determining when the Markov chain has converged to the point that it is generating values that can be considered as draws from the desired posterior distribution. Since in most animal models convergence for more than one parameter must be monitored for the Bayesian method, the multi-dimensional potential scale reduction (MPSR) (Brooks and Gelman, 1998) can be used to simultaneously diagnose the convergence of all parameters of interest. Nevertheless, it may take a long time for the MCMC algorithm to converge. It is therefore of interest to compare a variety of MCMC algorithms. This is the goal of this thesis.

A number of factors can affect convergence rate of MCMC algorithms. Alternative parameterizations of models may impact the convergence rate. Two reparameterizations are considered in this thesis to improve convergence rate: (1) use of orthogonal polynomials in place of regular polynomials, and (2) hierarchical centering. The former is motivated by the observation that the convergence rate for the MCMC algorithm de-

pends on the correlation of successive draws in a Markov chain simulation (Gilks and Roberts, 1996). The latter is based on the work of Gelfand et al. (1995), which shows that hierarchical centering can lead to improved convergence by changing the shape of the surface of the sampling distribution of parameters. Drawing parameters in "batches" rather than element-by-element can also affect convergence rate (Gilks et al., 1996).

The main objective of this study is to compare the performance of various MCMC algorithms in the Bayesian approach to analyzing longitudinal data. Chapter 2 reviews models that have been used for the analysis of longitudinal data in animal science, with specific emphasis on the linear and nonlinear random regression models that are the focus of this thesis. Inferential approaches to analyzing data with such models are reviewed in Chapter 3, emphasizing the Bayesian approach. Chapter 4 introduces the various techniques that can be used to improve the rate of convergence of MCMC methods. In Chapter 5 these techniques are applied to linear random polynomial regression models. A pig weight data set is used to evaluate the improvement in convergence affected by the various techniques for linear random regression models. Approaches for improving the convergence rate of MCMC in nonlinear random regression models are described in Chapter 6; these approaches are also applied to the pig weight data. Final conclusions are discussed in Chapter 7.

CHAPTER 2 MODELS FOR LONGITUDINAL DATA

2.1 Introduction

When the progress of a trait (e.g., weight) over time is of interest, one typically collects repeated measurements on a number of individuals. For longitudinal data of this type in animal populations, measurements of related animals are correlated and measurements across times for a single animal are also correlated. These correlations can be due to genetic effects (e.g., additive, dominance, and epistasis), permanent environmental effects (e.g., nutrition and climate), the interaction between genetic and environmental effects, and temporary environmental effects (e.g., measurement error and localized circumstances) (Falconer and Mackay, 1996).

We begin by considering a single time point. A simple model, the animal model, for a single phenotypic record of animal i is (Henderson, 1984; Van Vleck,1993)

$$y_i = \mathbf{x}_i\mathbf{b} + u_i + e_i,$$

where $\mathbf{x}_i$ is the row vector associated with the fixed effects $\mathbf{b}$ for animal i, u_i is the animal's additive genetic effect, and e_i is the residual. The residual incorporates permanent environmental effects, non-additive genetic effects, and any other factors unaccounted for in the other terms. Across a population of n animals with additive genetic relationship matrix $\mathbf{A}$ $(n \times n)$ (Wright, 1922), it is common to assume that u_i: $i = 1, \ldots, n$, are jointly normal, with $\mathbf{u} = (u_1 \ \ldots \ u_n) \sim N(0, \sigma_a^2\mathbf{A})$ and e_i: $i = 1, \ldots, n$ are iid $N(0, \sigma_e^2)$. It follows that the marginal distribution of y_i is $N(\mathbf{x}_i\mathbf{b}, \sigma_y^2 = \sigma_a^2 + \sigma_e^2)$. The ratio σ_a^2/σ_y^2

expresses the extent to which animals' phenotypes are determined by additive genetic effects and is called heritability (Falconer and Mackay, 1996).

When repeated measurements on the same characteristic are recorded on an animal, the between-animal variance can be partitioned as genetic and permanent environmental variances. The model for the j^{th} record of animal i is (Henderson, 1984; Van Vleck, 1993)

$$y_{ij} = \mathbf{x}_i\mathbf{b} + u_i + p_i + \epsilon_{ij},$$

where p_i is added to address the permanent environmental effects associated with all records of animal i, and ϵ_{ij} is the within-animal residual. Under the assumption of normality for all effects, we have $u_i \sim N(0, \sigma_a^2)$, $p_i \sim N(0, \sigma_p^2)$, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. When the relationship between animals is unknown or when animals are unrelated, u_i and p_i are confounded. When there is a single measurement for each animal, p_i and ϵ_i are confounded.

Let $\mathbf{y}_i$ be a vector containing all of the measurements for animal i. If we collect all of the repeated measurement data in a single vector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) = (y_{11}, y_{12}, \ldots)$, then the basic animal model used for longitudinal data can be written in the form

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \boldsymbol{\epsilon}. \tag{2.1}$$

with fixed effects $\mathbf{b}$, animal genetic effects $\mathbf{u}$, permanent environmental variation $\mathbf{p}$, and $\boldsymbol{\epsilon}$ representing temporary environmental variation (measurement error and other variation). The matrices $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{W}$ identify the various effects. They are different for different models, so we don't discuss them further now but will illustrate their structure in subsequent sections.

The remainder of this chapter reviews a number of models that have been proposed for the analysis of longitudinal data. Some major linear models used for longitudinal data are described in Sections 2.2 to 2.4. Covariance function models are discussed in Section 2.5; these are related to the random regression model of Section 2.4. Nonlinear random

regression models are discussed in Section 2.6. For convenience, common notation is described here, before introducing the models.

Notation

n: The total number of animals in the data set

r_i: The number of repeated records on animal i

N: The total number of records. $N=\sum_{i=1}^{n} r_i$

$\mathbf{A}$: The additive genetic relationship matrix between animals

$\mathbf{y}_i = (y_{i1}\ y_{i2}\ \ldots\ y_{ir_i})^T$: The observation vector for animal i

$\mathbf{y} = (\mathbf{y}_1^T\ \mathbf{y}_2^T\ \ldots\ \mathbf{y}_n^T)^T = (y_{11}\ y_{12}\ \ldots\ y_{nn_n})^T$: The entire observation vector sorted by animal.

$\mathbf{X}_i$: The design matrix associated with fixed effects for animal i.

$\mathbf{Z}_i$: The incidence matrix associated with the animal random effects or the additive genetic effects for animal i.

$\mathbf{W}_i$: The incidence matrix associated with the permanent environmental effects for animal i.

$\mathbf{X}$: The design matrix associated with fixed effects for all animals.

$\mathbf{Z}$: The incidence matrix associated with animal random effects or animal additive genetic effects for all animals.

$\mathbf{W}$: The incidence matrix associated with permanent environmental effects for all animals .

$\mathbf{b}_k$: The vector of the fixed effect parameters for the k^{th} subpopulation defined by the level of the fixed effects.

b: The vector of the fixed effect parameters for all animals.

u: The vector of the animal random effect parameters or additive genetic random effects for all animals.

p: The vector of the permanent environmental effects for all animals.

$\mathbf{1}_t$: A column vector with 1 for every entry

$\mathbf{I}_p$: A $p \times p$ identity matrix

$\mathbf{J}_p$: A $p \times p$ matrix with 1 for every entry

ϵ: The $N \times 1$ vector of random residuals, which are iid $N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$

σ_e^2: The variance for the random residuals

G: The variance-covariance matrix for the animal random effects or for additive genetic effects

E: The variance-covariance matrix for the permanent environmental effects

2.2 Repeatability models

In the repeatability model (Henderson, 1984; Littell et al., 1996), with t common measuring times for the n animals, the time effects are assumed common for all animals and treated as random effects. Letting $\mathbf{p} = (p_1 \ \ldots \ p_t)$ denote the time effects, we have

$$
\begin{aligned}
\mathbf{y} &= \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_n^T \end{pmatrix}^T \\
&= \begin{pmatrix} bf.X_1 \\ bf.X_2 \\ \vdots \\ bf.X_n \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{1}_t & 0 & \cdots & 0 \\ 0 & \mathbf{1}_t & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_t \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} + \begin{pmatrix} \mathbf{I}_t \\ \mathbf{I}_t \\ \vdots \\ \mathbf{I}_t \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_t \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \\
&= \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \boldsymbol{\epsilon}
\end{aligned}
$$

where $\mathbf{X}_i$, $\mathbf{I}_t$, $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{p}$ are described in the notation portion of Section 2.1. Note that the size of $\mathbf{y}_i$ is $t \times 1$, $\mathbf{y}$ is $nt \times 1$, $\mathbf{Z}$ is $nt \times n$, $\mathbf{u}$ is $n \times 1$, $\mathbf{W}$ is $nt \times t$, and $\mathbf{p}$ is $t \times 1$. The incidence matrix $\mathbf{Z} = [z_{ji}]$ $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, nt)$ is constructed with $z_{ji} = 1$ if the j^{th} record is taken from animal i and $z_{ji} = 0$ otherwise. The incidence matrix $\mathbf{W} = [w_{jk}]$ $(j = 1, 2, \ldots, nt;\ k = 1, 2, \ldots, t)$ is constructed with $w_{jk} = 1$ if the j^{th} record is taken at time k and with $w_{jk} = 0$ otherwise. The genetic effects, $\mathbf{u}$, are $N(0, \sigma_a^2\mathbf{A})$ or, if animals are unrelated, $N(0, \sigma_a^2\mathbf{I})$. For this model $\mathbf{p} \sim N(0, S(\sigma_t^2))$, where S is a pre-determined function of variance components (σ_t^2).

For the repeatability model, the animal genetic variations are assumed to be constant over time. Also, the heritability (σ_a^2/σ_y^2) is assumed to be constant over time, when the repeatability model is adopted. With respect to the structural correlation of repeated measurements, $S(\sigma_t^2)$, there are several popular choices (Littell et al., 1993). One example is compound symmetry, where the covariance matrix of $\mathbf{p}$ is S=$((\mathbf{J}_t\rho + \mathbf{I}_t(1-\rho))\sigma_t^2)$; where ρ is the correlation between any two observations on the same animal. Another example is the first-order autoregressive model (AR(1)), where the degree of correlation between two repeated records decreases when the time interval between them increases, so that S=$\left(\mathbf{K}\sigma_p^2\right)$ with $\mathbf{K} = [k_{ij}] = [\rho^{|i-j|}]$ $(1 \leq i, j \leq t)$.

Repeatability models are often used because they are easy to specify and because few parameters are required to accommodate the repeated measurements; e.g., two parameters for the compound symmetry model. One difficulty with the repeatability model is that it requires a prior assumption about the covariance structure of $\mathbf{p}$.

2.3 Multivariate mixed models

If the t measurements taken at different fixed times are considered as different traits, then multivariate mixed (MVM) models (Henderson, 1984; van Vleck and Boldman, 1993) may be used. A MVM model is similar to (2.1), but, for convenience, the records

in $\mathbf{y} = (\mathbf{y}_1 \ldots \mathbf{y}_j \ldots \mathbf{y}_t)^T$ are now arranged by time (trait) instead of by animal (i.e., $\mathbf{y}_j$ is an $n \times 1$ vector of the observations on all animals taken at time j). The MVM includes t such $n \times 1$ response vectors and separate vectors of fixed effects $\mathbf{b}_j$, genetic effects $\mathbf{u}_j$, and permanent environmental effects $\mathbf{p}_j$ for each measuring time. Let $\mathbf{G} = [G_{ij}]$ and $\mathbf{E} = [E_{ij}]$ denote $t \times t$ covariance matrices for genetic effects and permanent environmental effects among the t times, respectively. The matrices $\mathbf{G}$ and $\mathbf{E}$ are used to introduce correlations among repeated measurements for an individual. Suppose balanced data of repeated measurements are taken on all n animals at equal times. Then an MVM model is written in the form

$$
\begin{aligned}
\mathbf{y} &= \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_t^T \end{pmatrix}^T \\
&= \begin{pmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \mathbf{X}_t \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_t \end{pmatrix} + \begin{pmatrix} \mathbf{I}_n & 0 & \cdots & 0 \\ 0 & \mathbf{I}_n & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_t \end{pmatrix} \\
&\quad + \begin{pmatrix} \mathbf{I}_n & 0 & \cdots & 0 \\ 0 & \mathbf{I}_n & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{pmatrix} \\
&= \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \boldsymbol{\epsilon}.
\end{aligned}
$$

where $\mathbf{X}_j$, $\mathbf{I_n}$, $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{p}$ are described in the notation of Section 2.1. $\mathbf{u} = (\mathbf{u}_1 \ldots \mathbf{u}_t)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A})$ and $\mathbf{p} = (\mathbf{p}_1 \ldots \mathbf{p}_t)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{E} \otimes \mathbf{I}_n)$.

For unbalanced data, where not all animals are measured at all times, the identity matrices on the diagonal of $\mathbf{Z}$ or $\mathbf{W}$ are replaced by matrices with rows corresponding to the missing times removed, but $\mathbf{u}$ and $\mathbf{p}$ remain of full dimension. Note that information about u_{ji} is still available even when y_{ji} is not observed because of information from related animals and other measurements on that animal.

MVM models allow genetic and environmental variations to be different from time to time because of the effects of **G** and **E**. There are several difficulties for the application of MVM models. First, well-defined genetic and environmental covariance matrices describing the variation between the repeated measurements are required. Second, estimation of **G** and **E** requires measurements be taken at fixed times. In practice, animals are often measured at irregular times, so categorizing the time of measurement in order to apply MVM models may involve bias in estimation of **G** or **E**. Lastly, the large number of variance parameters involved in **G** and **E** may result in overparameterization and can cause problems in computation.

If no data are missing, a canonical transformation can be used to convert a multivariate analysis to a set of single-trait analyses, which significantly reduces the computational requirement (Lin and Smith, 1990). If missing data are present, the EM (expectation and maximization) method can be used to estimate the missing data before implementing the canonical transformation (van der Werf et al., 1998).

2.4 Random regression models

For longitudinal data such as lactation or growth data, the function describing the change in the response variable across time is critical for prediction. The models discussed thus far can be used for such data, but they are not designed to handle data observed at irregular times. Random regression (RR) models, which are also called multi-stage random effects models (Laird and Ware, 1982), can be adopted to extend animal models to address longitudinal data. RR models accommodate measurements at arbitrary times, allow for between-animal variation, and also reduce the number of variance components compared with multivariate models. In general, simple RR models can be set up in two stages: the first stage links the phenotypic traits to individual animal effects and within individual variation; in the second stage, between-individual

variation is modeled. The common effects for every individual can be thought of as population parameters, which are the fixed effects in the model, and the individual-specific coefficients, the random effects, explain the individual deviation from the population average. RR models are appealing in that every individual has its own model with population-level parameters estimated by averaging across individuals.

Let **Mpq** denote a model in which the response variable is fitted by polynomials of time, with p^{th} and q^{th} degree polynomials for the fixed and the random effects, respectively. Model M42, which includes quartic and quadratic polynomials for the fixed effects and the random effects, respectively, is used as an example for illustrating the setup for RR models. To begin, suppose that a response for animal i at time j is modeled by a quartic polynomial of time.

$$\begin{aligned} y_{ij} &= \beta_0 + t_{ij}\beta_1 + t_{ij}^2\beta_2 + t_{ij}^3\beta_3 + t_{ij}^4\beta_4 + \epsilon_{ij} \\ &= \mathbf{t^T b} + \epsilon_{ij}. \end{aligned}$$

where $\mathbf{t} = (1\ t_{ij}\ t_{ij}^2\ t_{ij}^3\ t_{ij}^4)^T$, $\mathbf{b} = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4)^T$, and $\epsilon_{ij} \sim N(\mathbf{0}, \sigma_e^2)$ represents the within animal variation. Note this quartic polynomial assumes the same coefficients (population parameters) apply to each animal. This is not likely to be a realistic assumption. The random regression model allows some or all of the coefficients to vary among individuals. In model M42, the first three $\beta = (\beta_0\ \beta_1\ \beta_2)$ are assumed to vary across animals; that is, we introduce additional variation $u_{ji}; j = 0, 1, 2$ which allows for the constant, linear, and quadratic terms to be specific to animal i. Placing the r_i records of animal i $(i = 1, \ldots, n)$ in a vector $\mathbf{y}_i$, the model for animal i can be written as

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ir_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & t_{i1}^3 & t_{i1}^4 \\ 1 & t_{i2} & t_{i2}^2 & t_{i2}^3 & t_{i2}^4 \\ & & \vdots & & \\ 1 & t_{ir_i} & t_{ir_i}^2 & t_{ir_i}^3 & t_{ir_i}^4 \end{pmatrix} \begin{pmatrix} \beta_0 + u_{0i} \\ \beta_1 + u_{1i} \\ \beta_2 + u_{2i} \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{ir_i} \end{pmatrix} \qquad (M42)$$

$$= \mathbf{T}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \epsilon_i, \tag{2.2}$$

where $\mathbf{T}_i$ is an $r_i \times 5$ incidence matrix, $\mathbf{b} = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4)^T$, $\mathbf{Z}_i$ is an $r_i \times 3$ incidence matrix made up of the first 3 columns of $\mathbf{T}_i$, and $\mathbf{u}_i = (u_{0i}\ u_{1i}\ u_{2i})^T$ is a vector of individual effects. This is known as the stage I model. In stage II, the random animal coefficients $\mathbf{u}_i$ are assigned a distribution. It is common to assume a Gaussian distribution, $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G})$, where $\mathbf{G}$ is a 3×3 covariance matrix for model M42.

In general, the population parameters may differ across m subpopulations, where subpopulations may be defined by year-season, gender, location, ...etc. With m subpopulations, suppose animals 1 and n belong to subpopulation 1, and animal 2 belongs to subpopulation 2, then the random regression model Mpq for all records is of the form

$$\begin{aligned}
\mathbf{y} &= \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_n^T \end{pmatrix}^T \\
&= \begin{pmatrix} \mathbf{T}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{T}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \mathbf{T}_n & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \\
&= \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \epsilon,
\end{aligned} \tag{2.3}$$

where $\mathbf{X}_i$ is a $r_i \times mp$ matrix with block matrix $\mathbf{T}_i$ placed in the columns corresponding to the subpopulation to which animal i belongs, $\mathbf{b}_k$ represents the $p+1$ coefficients of the p^{th} degree polynomial for subpopulation k, $\mathbf{u}_i$ is the $q+1$ animal-specific coefficients of the q^{th} degree polynomial for animal i with $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G})$, and random residuals $\epsilon \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$. Note that $\mathbf{G}$ is a $(q+1) \times (q+1)$ covariance matrix. Since animals are assumed to

be uncorrelated for the moment (i.e., genetic relationships are being ignored), it follows that $var(\mathbf{y}) = \mathbf{Z}var(\mathbf{u})\mathbf{Z}^T + var(\epsilon) = \mathbf{Z}(\mathbf{I} \otimes \mathbf{G})\mathbf{Z}^T + \sigma_e^2\mathbf{I}$. If permanent environment is another factor producing variation in the polynomial coefficients, then $\mathbf{Wp}$ may be added to (2.3). A more complete discussion of such models, that incorporates additive genetic relationships and permanent environmental variation, is provided in Chapter 5 of the thesis.

The dimensions of $\mathbf{G}$ (and $\mathbf{E}$, if $\mathbf{A}$ is incorporated) in random regression models are often much fewer than those of the corresponding matrices in the multivariate mixed models when the number of measuring times is large ($p + 1 \ll t$). In that case the number of parameters in the model, both location parameters and variance components, is reduced, which significantly reduces the computational burden. Moreover, the RR model can be applied to data measured at irregular times, since the time effects on the response variable are modeled directly. There is no requirement that measurements are taken at the same time.

The order of the polynomials used for the fixed or random effects can perhaps be determined with likelihood ratio tests (Efron, 1967). An alternative way to determine the order of fit for the random effects is using the covariance function (CF) models of Kirkpatrick and Heckman (1989). CF models directly model the covariance between observations taken on the same individual at two different times rather than modeling the underlying trait measurements. Suppose there are at most s repeated measurements on an animal and measuring times are rescaled to be within [-1, 1] in order to satisfy the domain of Legendre polynomial functions (see Section 5.4 for more details). Then a $s \times s$ covariance matrix estimated by s repeated measures (e.g., genetic covariance matrix or permanent environmental covariance matrix), say $\mathbf{V}$, can be expressed in terms of orthogonal functions as (Kirkpatrick et al., 1990; Kirkpatrick et al., 1994)

$$\mathbf{V} = \mathbf{\Phi}\mathbf{C}_k\mathbf{\Phi}^T + \mathbf{D},$$

where $\mathbf{\Phi}$ is now a $s \times k$ $(s \geq k)$ matrix of Legendre polynomials and $\mathbf{D}$ is a matrix with errors due to the discrepancy between elements of $\mathbf{V}$ and estimates obtained from fitting the k^{th} degree CF (with $\mathbf{D} = \mathbf{0}$ when $s = k$). Note that once a CF is modeled the covariance at any time point, even those not in the original $\mathbf{V}$ can be estimated. The covariance between measures at any two times, t_i and t_j, can then be estimated as

$$\hat{V}(i,j) = \phi(t_i)^T \mathbf{C}_k \phi(t_j).$$

Meyer (1998) has shown that CF models are equivalent to random regression models and also shows that the entries of the matrix of coefficients $\mathbf{C}_k$ are equal to the covariances between the random regression coefficients if k^{th} degree Legendre polynomial are used in the RR model.

2.5 Nonlinear models

Random polynomial regression models are linear in the parameters and are easy to fit, but in some situations such linear models do not accurately reflect the biological process being modeled. For example, a limit to growth is often observed for animal growth curves, i.e., the growth curve flattens out. Polynomial models are not able to fit such features well, although they may be an adequate approximation if the data are collected over a limited range of the growth cycle. Nonlinear models are an alternative approach. A nonlinear function is defined as a function in which at least one of its parameters appears nonlinearly. In a formal sense, nonlinear means that at least one of the derivatives of the response variable with respect to the functional parameters is a function of at least one of those parameters (Ratkowsky, 1990). Nonlinear functions can be adopted to model the response in terms of biologically meaningful parameters.

We assume that the j^{th} observed response on animal i occurs at day t_{ij} and can be expressed as the sum of a non-linear function $f(.)$ depending on parameter $\boldsymbol{\theta}$ and

random variation

$$y_{ij} = f(\boldsymbol{\theta}, t_{ij}) + \epsilon_{ij}$$

where ϵ_{ij} is a random residual with $\epsilon_{ij} \sim N(0, \sigma_e^2)$.

As with RR polynomial models, it is often desirable to allow each animal to have its own underlying growth curve. Subject-specific effects are introduced into the model by allowing $\boldsymbol{\theta}$ to vary from individual to individual. Lindstrom and Bates (1990) introduce the **nonlinear mixed (effects) models** (NLM models) by generalizing the linear mixed model and the standard fixed effects nonlinear model. The NLM models allow the parameters of the nonlinear function to be affected by fixed effects and random effects that are associated with individuals; that is,

$$y_{ij} = f(\boldsymbol{\theta}_i, t_{ij}) + \epsilon_{ij}, \qquad \boldsymbol{\theta}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i. \tag{2.4}$$

where $\mathbf{X}_i$ is the incidence matrix associated with the fixed effects $\mathbf{b}$ for individual i, $\mathbf{Z}_i$ is the incidence matrix for the random effects $\mathbf{u}_i$ of individual i. Since NLM models can also be set up by a two-stage procedure, we can regard nonlinear mixed effects models as one class of random regression models.

The choice of which nonlinear functions to use is in practice data-dependent. In general, when the response variable is a growth trait, the graph of growth response against time has a sigmoidally shaped curve. Several one-parameter to four-parameter nonlinear functions for sigmoidally shaped curve were reviewed by Ratkowsky (1990). Features of the functions can include an upper asymptote, a growth rate parameter, an inflection point, and a lower asymptote as the number of parameters increases. If the responses have a lower asymptote of zero and a finite upper asymptote, then the **logistic model** can be used to fit the growth curve. The logistic model takes

$$y_{ij} = \alpha_i / \left(1 + exp(\beta_i - \frac{t_{ij}}{\kappa_i})\right) + \epsilon_{ij},$$

where t_{ij} is the time of the j^{th} measurement for individual i, $\boldsymbol{\theta}_i = (\alpha_i, \beta_i, \kappa_i)^T$ is the parameter vector for individual i, α_i is the upper asymptotic value of the response, β_i

is the growth rate, and κ_i is the inflection point of the growth rate. The logistic curve is skew-symmetric, with an inflection point at $t_{ij} = \beta_i\kappa_i$, where $E(y_{ij} \mid t_{ij}, \boldsymbol{\theta}_i) = \alpha_i/2$. Figure 2.1 shows the mean curve of the logistic model for selected parameter values.

When the sigmoidal curve is asymmetric about its inflection point, the Gompertz function can be used to model the curve. The **Gompertz function** model is

$$y_{ij} = \alpha_i exp\left(-exp\left(-\beta_i\left(\frac{t_{ij}}{\kappa_i} - 1\right)\right)\right) + \epsilon_{ij}$$

where t_{ij} and $\boldsymbol{\theta}_i = (\alpha_i, \beta_i, \kappa_i)^T$ are of the same meaning as those in logistic model. Figure 2.2 shows the Gompertz curve for selected parameter values.

The nonlinear models, especially the Gompertz models, are discussed more fully in Chapters 4 and 6.

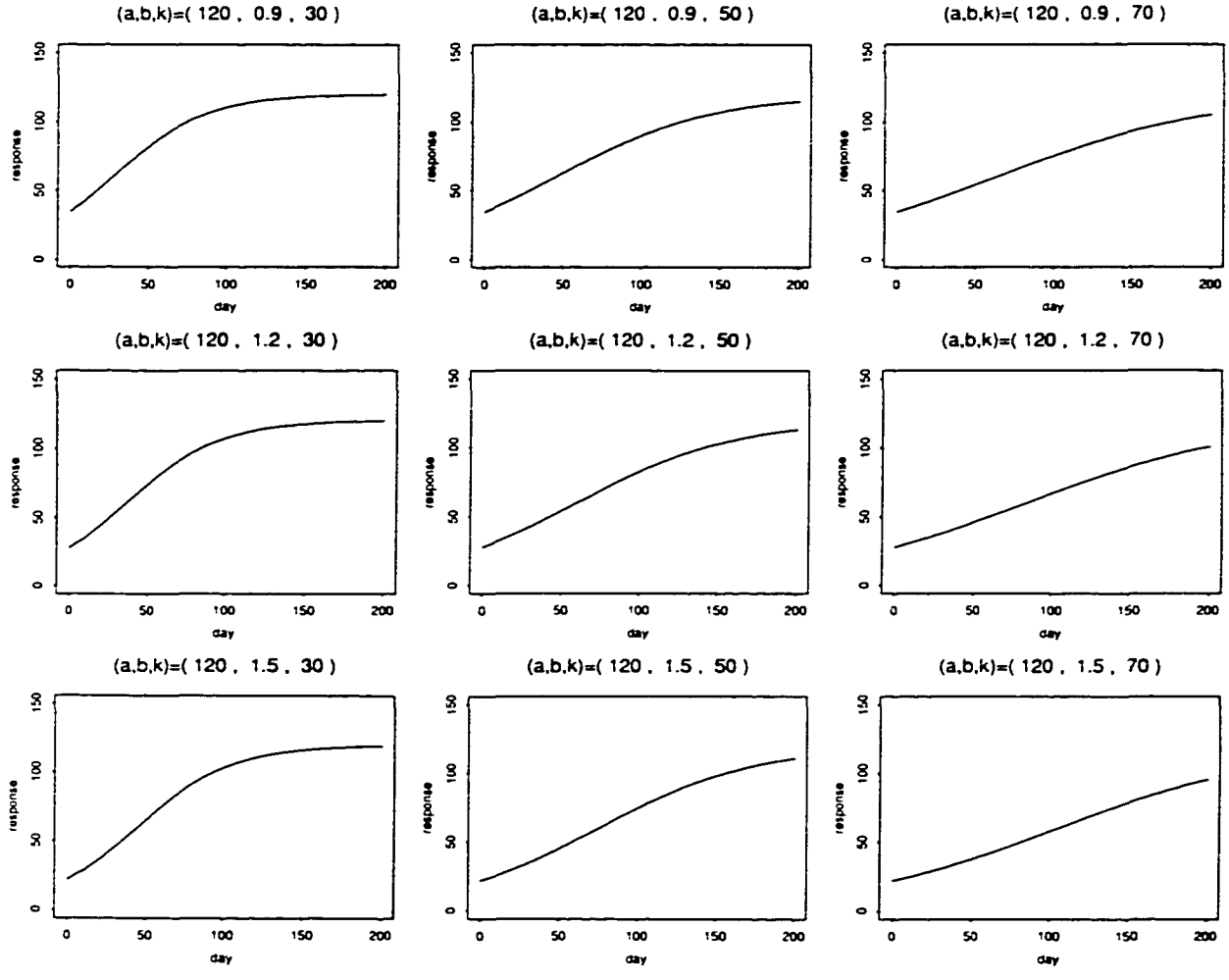

Figure 2.1 Logistic curves for selected parameter values.

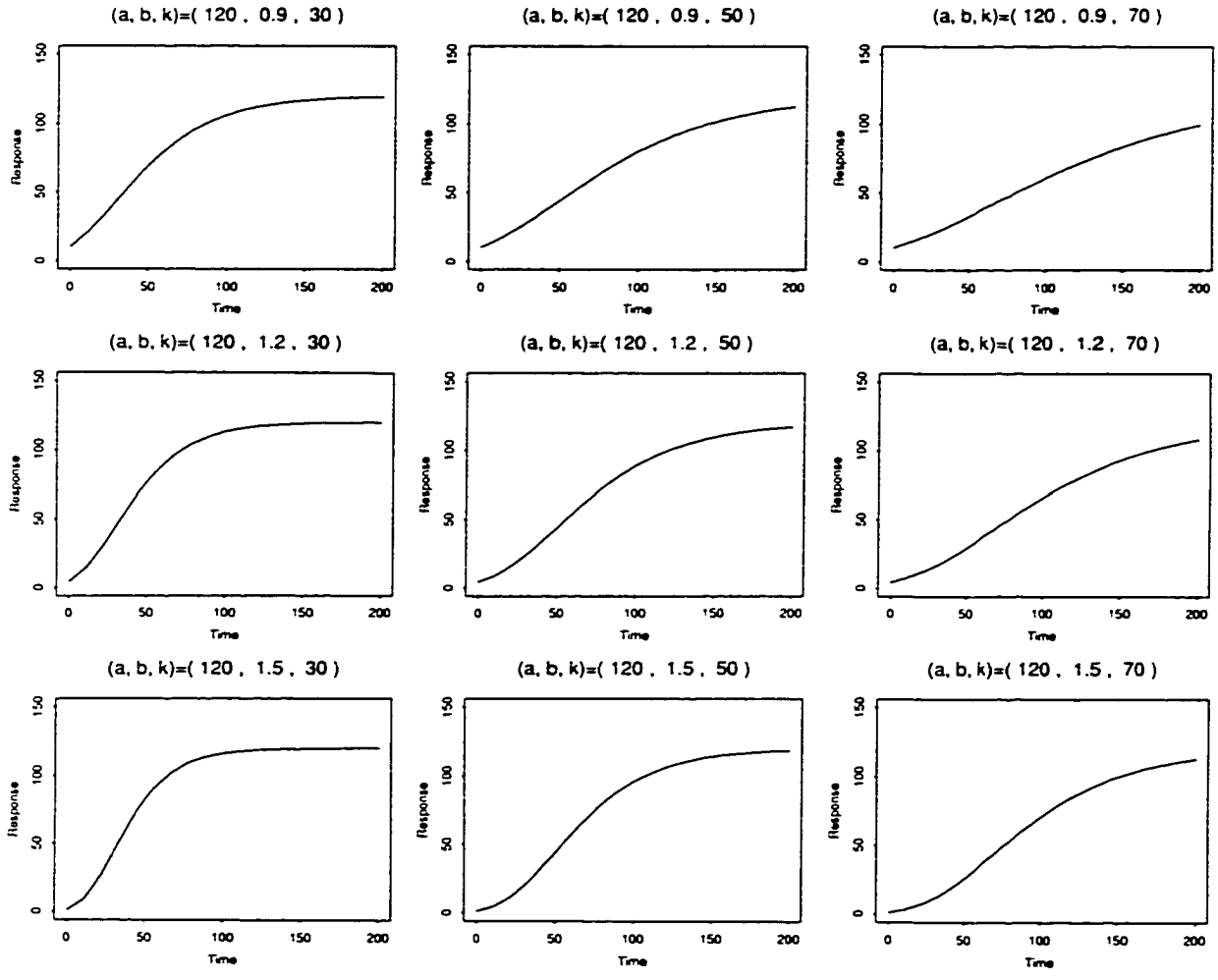

Figure 2.2 Gompertz curves for selected parameter values.

CHAPTER 3 STATISTICAL INFERENCE

3.1 Introduction

When a statistical model is set up to describe a data set, making inferences for model parameters and predictions for future data are of great interest. There are two main approaches for drawing inferences for model parameters: likelihood-based inference (i.e., maximum likelihood estimation and associated inference) and Bayesian inference.

Let $\mathbf{y}$ denote a vector of observable quantities and $\boldsymbol{\theta}$ the vector of model parameters. A statistical model is a joint distribution for $\mathbf{y}$ given the set of parameters $\boldsymbol{\theta}$, $p(\mathbf{y} \mid \boldsymbol{\theta})$. If this distribution is viewed as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}$, then it is known as the likelihood function, $L(\boldsymbol{\theta} \mid \mathbf{y})$. Maximum likelihood estimates are obtained by finding the value for $\boldsymbol{\theta}$ such that $L(\boldsymbol{\theta} \mid \mathbf{y})$ is maximized. Typically inference is based on asymptotic properties of the maximum likelihood estimates: the estimates have good asymptotic repeated sampling properties (Miller, 1973).

Bayesian inference is so named because of it relies on Bayes' rule (noted by Rev. Bayes in 1763), concerning conditional probabilities. Bayes' rule states

$$\text{p(hypothesis | datum)} = \frac{\text{p(datum | hypothesis) p(hypothesis)}}{\text{p(datum)}}$$

In this case, hypothesis can refer to a model or to a true state of nature regarding the parameter values. Under the Bayesian perspective, observations and parameters of a statistical model are both considered as random variables. Model specification begins with a joint distribution for $\mathbf{y}$ and $\boldsymbol{\theta}$, $p(\mathbf{y}, \boldsymbol{\theta})$. The joint distribution is typically

specified as product of $p(\mathbf{y} \mid \boldsymbol{\theta})$, the sampling distribution for $\mathbf{y}$ given $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta})$, a prior distribution for the parameter $\boldsymbol{\theta}$. The prior distribution provides an mechanism for incorporating uncertainty about $\boldsymbol{\theta}$. Inference for $\boldsymbol{\theta}$ given observed data $\mathbf{y}$ follows directly from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$, which by Bayes' rule is a conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$ (see Section 3.3).

In this chapter, likelihood-based inference and Bayesian inference approaches are reviewed in Sections 3.2 and 3.3, respectively. Markov chain Monte Carlo (MCMC) methods, which are simulation-based approaches to Bayesian inference, and the convergence criteria used for terminating the simulations will also be described in Section 3.3. Some comments on the relationship of the Bayesian approach and the likelihood-based approach are given in Section 3.4.

The model considered in this chapter assumes that $\mathbf{y}$ is continuous and follows a Gaussian distribution. The general form of the model is

$$\mathbf{y} = f(\boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{3.1}$$

where $\boldsymbol{\theta}$ is parameter vector, $f(.)$ can be either a linear or nonlinear function of covariates $\mathbf{X}$ and parameter vector $\boldsymbol{\theta}$, and sampling errors $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$. The error variance matrix $\mathbf{R}$ is often assumed to be of simple form such as $\mathbf{R} = \mathbf{I}\sigma_e^2$ (independence among residuals). For the linear mixed model $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{u})$ and $f(\boldsymbol{\theta}) = \mathbf{Xb} + \mathbf{Zu}$ with $\mathbf{X}$ and $\mathbf{Z}$ incidence matrices associated with the fixed effects $\mathbf{b}$ and the random effects $\mathbf{u}$. The random effects $\mathbf{u} = (\mathbf{u}_1 \ \ldots \ \mathbf{u}_n) \sim N(\mathbf{0}, \mathbf{A} \otimes \mathbf{G})$ introduce individual variation from the population average $\mathbf{Xb}$, where $\mathbf{A}$ is the additive genetic relationship matrix for the n animals and $\mathbf{G}$ is the variance matrix of individual random effects $\mathbf{u}_i$. As described in Chapter 2, we can included permanent environmental effects in the model when the relationship between animals is incorporated, and this will modify the model. For the purpose of this chapter, the simple model without permanent environmental effects is used.

3.2 Likelihood-based inference

If observations y_i, $(i = 1, \ldots, n)$ are independent, then the joint distribution or likelihood function is

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_i p(y_i \mid \boldsymbol{\theta}).$$

The likelihood-based approach to inference for $\boldsymbol{\theta}$ relies on this likelihood function, and inference is typically based on the maximized likelihood (ML) estimates of $\boldsymbol{\theta}$.

3.2.1 Maximum likelihood estimation

Denote the log-likelihood function as $l = log(L(\boldsymbol{\theta} \mid \mathbf{y}))$. The MLE of $\boldsymbol{\theta}$ satisfies $\frac{\partial l}{\partial \boldsymbol{\theta}} = 0$ and $\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ is negative definite at the MLE. The MLE can be obtained via a number of optimization methods (Thisted, 1988). For example, the MLE for $\boldsymbol{\theta}$ can be obtained by the Newton-Raphson method; in that iterative algorithm the value at iteration $t-1$ is updated by

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \mathbf{H}^{-1}\mathbf{F},$$

where $\mathbf{F}$ and $\mathbf{H}$ are the first and the second derivative with respect to $\boldsymbol{\theta}$ of the log-likelihood function evaluated at $\boldsymbol{\theta}^{t-1}$ and are called the gradient vector and the Hessian matrix, respectively.

Although ML estimators are not generally unbiased, they have several appealing features. The ML estimators are functions of sufficient statistics, they are efficient and they are consistent. MLE are asymptotically normal with mean equal to the true parameter value and variance matrix equal to the inverse of the Fisher information, which is the expectation value of the negative of $\mathbf{H}$ (Miller, 1973; Searle et al., 1992).

3.2.2 Role of maximum likelihood in linear mixed models

Consider the simple linear mixed model,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\epsilon \sim N(\mathbf{0}, \mathbf{R})$ and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A} \otimes \mathbf{G})$. Assume the joint distribution of $\mathbf{y}$ and $\mathbf{u}$ is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{Xb} + \mathbf{Zu} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \otimes \mathbf{G} \end{pmatrix} \right)$$

Strictly speaking, one can not apply the maximum likelihood approach here because the $\mathbf{u}$ are not observed data. However, in practice estimates of $\mathbf{b}$ and predictions of $\mathbf{u}$ for given values of variance parameters are obtained by maximizing the joint distribution of $\mathbf{b}$ and $\mathbf{u}$. The logarithm of the joint density is

$$\begin{aligned} l_0 = -2ln L_0 &= \text{constant} + log|\mathbf{R}| + log\left|\mathbf{A} \otimes \mathbf{G}\right| \\ &\quad + (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) + \mathbf{u}^T (\mathbf{A}^{-1} \otimes \mathbf{G}^{-1}) \mathbf{u}. \end{aligned} \tag{3.3}$$

Let ω contain the unique variance components in $\mathbf{G}$ and $\mathbf{R}$. Given ω, maximization of the joint distribution of $\mathbf{y}$ and $\mathbf{u}$ with respect to $\mathbf{b}$ and $\mathbf{u}$ gives Henderson's mixed model equations (MME) (Henderson, 1950),

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + (\mathbf{A}^{-1} \otimes \mathbf{G}^{-1}) \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Y} \end{pmatrix}. \tag{3.4}$$

Thus, given variance components ω, an estimate of $\mathbf{b}$ and optimal prediction of $\mathbf{u}$ in (3.2) can be obtained by solving Henderson's MME.

An alternative way to find an estimate of $\mathbf{b}$ and prediction of $\mathbf{u}$ is described by Harville (1977). First, an estimate of $\mathbf{b}$ is obtained by maximizing the marginal distribution of $\mathbf{y}$ (averaged over the random effects $\mathbf{u}$).

$$\mathbf{y} \mid \mathbf{b}, \mathbf{R} \sim N\left(\mathbf{Xb}, \mathbf{V} = \mathbf{Z}(\mathbf{A} \otimes \mathbf{G})\mathbf{Z}^T + \mathbf{R}\right).$$

Its log-likelihood function l is

$$l = -2ln\ \mathbf{L} = \text{constant} + log|\mathbf{V}| + (\mathbf{y} - \mathbf{Xb})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{Xb}). \tag{3.5}$$

Then, given the estimates of the variance components, the MLE of $\mathbf{b}$ in this marginal distribution can be any solution to normal equations $(\mathbf{X^T V^{-1} X})\hat{\mathbf{b}} = \mathbf{X^T V^{-1} y}$. Then,

the optimal predictor of $\mathbf{u}$ can be obtained from its distribution conditional on $\mathbf{y}$.

$$\hat{\mathbf{u}} = (\mathbf{A} \bigotimes \mathbf{G})\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = (\mathbf{I} + \mathbf{ZSZ}(\mathbf{A} \bigotimes \mathbf{G}))^{-1}\mathbf{Z}^{\mathbf{T}}\mathbf{Sy}.$$

where $\mathbf{S} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{X})^{-}\mathbf{X}\mathbf{R}^{-1}$ and $(.)^{-}$ denotes a generalized inverse (Harville, 1976). Note that when $\mathbf{R}$ has a simple form, it is beneficial to find $\mathbf{u}$ in terms of $\mathbf{R}^{-1}$.

The random variables $\hat{\mathbf{u}}$ (that we have called optimal predictors) are in fact best linear unbiased predictors (BLUP) (Goldberger, 1962; Henderson, 1963), since they are linear functions of data, unbiased, best in the sense of minimum mean squared error, and predictors to distinguish them from estimators of the fixed effects. Hence, for a linear combination $\boldsymbol{\lambda}_1^T\mathbf{b} + \boldsymbol{\lambda}_2^T\mathbf{u}_i$, the BLUP is $\boldsymbol{\lambda}_1^T\hat{\mathbf{b}} + \boldsymbol{\lambda}_2^T\hat{\mathbf{u}}_i$, provided that $\boldsymbol{\omega}$ is known and $\boldsymbol{\lambda}_1^T\mathbf{b}$ is estimable (Henderson, 1975; Harville, 1976). BLUP estimates of animal breeding values are used in selection (Henderson, 1950; Harville, 1977; Robinson, 1991).

Let $\mathbf{C}$ be the coefficient matrix in Henderson's MME. The inverse of $\mathbf{C}$ can be partitioned corresponding to $(\mathbf{b}, \mathbf{u})$ as

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{pmatrix}.$$

The variances for the estimates and predictions are $var(\hat{\mathbf{b}}) = \mathbf{C}^{11}$ and $var(\hat{\mathbf{u}}) = \mathbf{A} \bigotimes \mathbf{G} - \mathbf{C}^{22}$. The variances for the bias of estimates for $\mathbf{u}_i$ are equal to the corresponding block matrix on the diagonal of $\mathbf{C}^{22}$, i.e., $var(\hat{\mathbf{u}}_i - \mathbf{u}_i) = \mathbf{C}_{ii}^{22}$. Inferences for $\mathbf{b}$ and $\mathbf{u}$ can be made according to asymptotic normal theory. For example, the estimated prediction variance is $var(\boldsymbol{\lambda}_1^T\hat{\mathbf{b}} + \boldsymbol{\lambda}_2^T\hat{\mathbf{u}}_i) = (\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)\mathbf{C}^{-1}(\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)^T$. A problem with such prediction variance is that they assume the variance components $\boldsymbol{\omega}$ are known. Adjustment is possible to account for uncertainty in the variance components (Harville, 1985).

3.2.3 Restricted maximum likelihood (REML) estimates

The previous discussion assumed the variance components $\boldsymbol{\omega}$ are known. When $\boldsymbol{\omega}$ is unknown, the MLE of $\boldsymbol{\omega}$ can be obtained from the marginal distribution of $\mathbf{y}$ (3.5). For simplicity, in this subsection, we focus on the case with $\mathbf{R} = \mathbf{I}\sigma_e^2$. The MLE of $\boldsymbol{\omega}$ is biased, since it does not take into account the loss in degrees of freedom from estimating fixed effects. Since the distribution of $\boldsymbol{\omega}$ does not depend on fixed effects, an alternative estimator for $\boldsymbol{\omega}$ is obtained from the distribution of so-called error contrasts, the likelihood function of which is independent of fixed effects (Patterson and Thompson, 1971; Harville, 1977). Suppose there are n_f independent columns in $\mathbf{X}$, which may be a non-full rank matrix, and let $\mathbf{X}^*$ be made up of any n_f linearly independent columns of $\mathbf{X}$. Define $\mathbf{B}$ as a $(N - n_f) \times N$ (N is the total number of observations) transformation matrix whose rows are any $N - n_f$ linear independent rows of the matrix

$$\mathbf{I} - \mathbf{X}^{\mathbf{T}}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-}\mathbf{X}^{\mathbf{T}} = \mathbf{I} - \mathbf{X}^{*\mathbf{T}}(\mathbf{X}^{*\mathbf{T}}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\mathbf{T}}$$

and then let $\mathbf{y}_1 = \mathbf{B}\mathbf{y}$. The elements of $\mathbf{y}_1$ are called error contrasts since $\mathrm{E}(\mathbf{y}_1)=\mathrm{E}(\mathbf{B}\mathbf{y}) = \mathbf{0}$. They are independent of $\mathbf{X}$, and the log-likelihood function associated with the error contrast $\mathbf{y}_1$ is

$$l_1 = -2ln\ \mathbf{L}_1 = constant + log|\mathbf{V}| + log\left|\mathbf{X}^{*T}\mathbf{V}^{-1}\mathbf{X}^{*}\right| + (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}), \quad (3.6)$$

where $\mathbf{V} = \mathbf{Z}(\mathbf{A} \otimes \mathbf{G})\mathbf{Z}^T + \mathbf{I}\sigma_e^2)$ and $\hat{\mathbf{b}}$ is a solution to $(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})\mathbf{b} = \mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{y}$. The restricted maximum likelihood (REML) estimators of the variance components are a set of solutions that maximize l_1. The asymptotic variance of REML estimates can be obtained from the inverse of the information matrix, which is the expectation of the negative of the Hessian matrix $\mathbf{H}$ with (i, j) entry equal to the second derivative of l_1 with respect to the i^{th} and the j^{th} element of $\mathbf{b}$ and $\boldsymbol{\omega}$ (Harville, 1975).

$$\mathbf{I}(\hat{\boldsymbol{\omega}}) = \left(-E\left(\frac{\partial^2 l}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\right)\right)^{-1}$$

Likelihood-based inference for linear mixed models is based on the REML estimates of variance components and BLUP estimates for the fixed coefficients and random effect coefficients. The algorithms used to carry out REML estimates will be discussed in the next subsection.

3.2.4 Algorithms used to compute REML estimates

Methods for identifying the REML estimates of variance components have been reviewed by Meyer and Smith (1996). The EM (expectation-maximization) algorithm, the DF-REML (derivative-free REML) algorithm, and the AI-REML (average information REML) algorithm are often used to compute REML estimates. We briefly describe these algorithms as follows.

The REML estimates of variance components can be derived using the expectation-maximization (EM) method (Lindstrom and Bates,1988; Searle et al., 1992), which is also called the nonlinear Gauss-Seidel method (Thisted, 1988). The EM-REML method first estimates variance components ω (E-step) by assigning initial values for the variance components, then in the maximization step (M-step) model coefficients are calculated by solving Henderson's MME as if ω is known. The E-step next re-estimates the variance components given the M-step results. The E-step and M-step alternate until convergence is reached. However, for large data sets the dimension of the coefficient matrix for the mixed model is large, and the time-consuming step of solving Henderson's MME can hinder the use of the EM-REML algorithm.

The DF-REML algorithm and the AI-REML algorithm have been developed to reduce the computational burden of REML estimation. The DF-REML algorithm was proposed by Graser et al. (1987) and has been expanded in use by Meyer (1989, 1991) and Boldman et al. (1993) via the well developed software MTDRREML (Boldman et al., 1993) and DFREML (Meyer, 1997).

In essence, the DF-REML algorithm obtains the inverse of the large matrix (e.g.,

$\mathbf{V}$) by Gaussian elimination of one row at a time, which makes the calculation of the last two terms in (3.6) more feasible. The strategy used for DF-REML algorithm is to evaluate the likelihood function without the calculation of the solution to MME, without the inverse of the coefficient matrix, and without the computation of any variance components. It first fixes all variance components but one, say σ_i^2. Then it evaluates (3.6) for four or more values of σ_i^2, and uses a quadratic approximation to find the value for σ_i^2 which maximizes (3.6). This step is repeated for each variance component in sequence until the REML estimates are found. The range for the four picked values is decreased as the process proceeds in order to get more accurate estimates for σ_i^2.

Now we briefly describe the strategy for AI-REML algorithm. Define

$$\mathbf{P} = \mathbf{V^{-1}} - \mathbf{V^{-1}X}(\mathbf{X}^T\mathbf{V^{-1}X})^{-1}\mathbf{X}^T\mathbf{V^{-1}},$$

then the last term in (3.6) can be written as $\mathbf{y}^T\mathbf{P}\mathbf{y}$ (Harville, 1977). The REML estimates for $\boldsymbol{\omega}$ require the first and second derivative with respect to the elements in $\boldsymbol{\omega}$. The Newton-Raphson method uses the negative of the observed Hessian matrix $-\mathbf{H}(\omega) = -[\mathbf{H}(\omega)_{ij}] = -[\partial^2 l_1/\partial\omega_i\partial\omega_j]$, while Fisher's method of scoring uses the expectation of the negative of the Hessian matrix, $\mathbf{I}(\omega)$=E($-\mathbf{H}(\omega)$). In both $-\mathbf{H}(\omega)$ and $\mathbf{I}(\omega)$ the traces of the matrix $\frac{\partial V}{\partial\omega_i}P\frac{\partial V}{\partial\omega_j}$ are involved, but with opposite signs.

$$-\mathbf{H}(\omega) = -\frac{\partial^2 l_1}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T} = -tr(\frac{\partial\mathbf{V}}{\partial\boldsymbol{\omega}}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\omega}}\mathbf{P}) + 2\mathbf{y}^t\mathbf{P}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\omega}}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\omega}}\mathbf{P}\mathbf{y},$$

$$\mathbf{I}(\omega) = E(-\mathbf{H}(\omega)) = tr(\frac{\partial\mathbf{V}}{\partial\boldsymbol{\omega}}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\omega}}\mathbf{P})$$

The average of $-\mathbf{H}(\omega)$ and $\mathbf{I}(\omega)$, say AI(ω), is only related to the first derivatives $\left[\frac{\partial V}{\partial\omega_i}\right]$, and consequently is a simple expression that is easy to compute. Hence, the AI-REML algorithm uses a Newton-type procedure by using AI (ω) in place of the second derivative matrix (Johnson and Thompson, 1995; Gilmour et al., 1995).

3.2.5 Estimation for nonlinear mixed models

All of the proceeding algorithms assume that we are working with a linear mixed model. When $f(\boldsymbol{\theta})$ is a nonlinear function in $\boldsymbol{\theta}$, determining estimates for $\boldsymbol{\theta}$ requires different technique. The Gauss-Newton algorithm and linear approximation are conventional techniques used for nonlinear models (Bates and Watts, 1988).

Assuming that the individual parameters $\boldsymbol{\theta}_i$ are determined by fixed effect parameters $\mathbf{b}$ and random effect parameters $\mathbf{u}_i$ and the expectation of $\mathbf{u}_i$ is $\mathbf{0}$, Sheiner and Beal (1980) expand $f(\boldsymbol{\theta})$ in a first-order Taylor series expansion of the random effects about the zero vector, so that $\mathbf{y}$ is approximated as $\mathbf{y} = f(\boldsymbol{\theta}) + \boldsymbol{\epsilon} \approx f(\boldsymbol{\theta}^* = (\mathbf{b}\ \mathbf{u} = \mathbf{0})) + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ with $\mathbf{Z} = \left.\frac{\partial f(\boldsymbol{\theta})}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{0}}$. If we make the approximation $f(\boldsymbol{\theta}^*) = \mathbf{X}_i\mathbf{b}$, then estimation can be done using the REML-BLUP method for linear mixed models.

Lindstrom and Bates (1990) introduce a likelihood-based method for inference in nonlinear mixed models when the normality assumption is made for both random effects and residuals. They work under the assumption that $\boldsymbol{\theta}_i$ can vary from individual to individual according to its associated factors; say $\boldsymbol{\theta}_i = \mathbf{A}_i\mathbf{b} + \mathbf{B}_i\mathbf{u}_i$, where $\mathbf{A}_i$ and $\mathbf{B}_i$ are incidence matrices associated with the fixed effects $\mathbf{b}$ and the random effects $\mathbf{u}_i$, respectively. The Lindstrom and Bates method (Lindstrom and Bates, 1990; Wolfinger, 1993) comprises a two-step iterative procedure: a pseudo-data step and a linear mixed effects step. In the pseudo-data step, a linear approximation to the likelihood function is derived by applying a first-order Taylor series expansion to the nonlinear function at the current value of the parameter vector $\boldsymbol{\theta}_i$ for individual i. Let $\boldsymbol{\theta}_i^*$ denote the current value for $\boldsymbol{\theta}_i$, $\mathbf{X}_i = \left.\frac{\partial f(\boldsymbol{\theta})}{\partial \mathbf{b}}\right|_{\mathbf{b}=\mathbf{b}^*}$ and $\mathbf{Z}_i = \left.\frac{\partial f(\boldsymbol{\theta})}{\partial \mathbf{u}_i}\right|_{\mathbf{u}_i=\mathbf{u}_i^*}$, then the pseudo-data step is derived as follows.

First we write

$$y_{ij} = f(t_{ij}, \boldsymbol{\theta}_i = \mathbf{A}_i\mathbf{b} + \mathbf{B}_i\mathbf{u}_i) + \epsilon_{ij}$$

$$\approx f(t_{ij}, \boldsymbol{\theta}_i^* = \mathbf{A}_i\mathbf{b}^* + \mathbf{B}_i\mathbf{u}_i^*) + \left.\frac{\partial f(\boldsymbol{\theta})}{\partial \mathbf{b}}\right|_{\mathbf{b}=\mathbf{b}^*} (\mathbf{b} - \mathbf{b}^*) + \left.\frac{\partial f(\boldsymbol{\theta})}{\partial \mathbf{u}_i}\right|_{\mathbf{u}_i=\mathbf{u}_i^*} (\mathbf{u}_i - \mathbf{u}_i^*) + \epsilon_{ij}.$$

then the pseudo data are

$$y_{ij}^* = y_{ij} - f(t_{ij}, \boldsymbol{\theta}_i^*) + \mathbf{X}_i\mathbf{b}^* + \mathbf{Z}_i\mathbf{u}_i^*,$$

and then model $y_{ij}^* = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \epsilon_{ij}$. Note that this final expression for the pseudo-data is in the form of the traditional linear mixed model. Then, in the linear mixed effects step, estimates for $\mathbf{b}$, $\mathbf{u}$ and variance components can be obtained by the methods mentioned in the previous section. The procedure is repeated until convergence of the estimates.

3.3 Bayesian inference

Bayesian inferences for quantities of interest are made in terms of probability statements conditional on the observed data. The most common situation is that in which the quantities of interest are model parameters, denoted by $\boldsymbol{\theta}$. The procedures for implementing the Bayesian approach are briefly summarized as follows.

1. Set up a probability model, this includes the sampling distribution $p(\mathbf{y} \mid \boldsymbol{\theta})$ and the prior distribution $p(\boldsymbol{\theta})$, where $\mathbf{y}$ are the observed data and $\boldsymbol{\theta}$ are the parameters involved in the likelihood function.
2. Derive the posterior distribution of the parameters, $p(\boldsymbol{\theta} \mid \mathbf{y})$. The joint distribution for $\boldsymbol{\theta}$ and $\mathbf{y}$ is $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. Bayes' Rule is used to calculate the posterior distribution of the parameters,

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{y}, \boldsymbol{\theta})/p(\mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}),$$

where the normalizing constant $p(\mathbf{y})$ is omitted, since it is independent of $\boldsymbol{\theta}$. The dimension of $\boldsymbol{\theta}$ is often large enough that analytic study of $p(\boldsymbol{\theta} \mid \mathbf{y})$ is not practical. In that case, we may adopt Markov chain Monte Carlo (MCMC) methods to

explore the joint posterior distribution (see, for example, Chib and Greenberg, 1995).

3. Evaluate the fit of the model.

For the purpose of our study, only the first two steps will be discussed in this section.

3.3.1 Model specification

3.3.1.1 Sampling distribution

The sampling distribution or the likelihood function relates the data $\mathbf{y}$ to the parameter $\boldsymbol{\theta}$. We assume that most quantitative variables in this thesis can be described by a normal distribution. The choice is analytically convenient. In some models it can be justified by the Central Limit Theorem, e.g., when a trait is presumed to reflect the sum of a large number of factors. As the choice of sampling distribution is common to both the Bayesian and likelihood-based approaches, we do not discuss it further here.

3.3.1.2 Prior distributions

The prior distribution describes uncertainty about the parameter values prior to collecting the data $\mathbf{y}$. There are two approaches to selecting prior distributions, the subjective and objective approaches (Berger, 1985). A subjective prior distribution is determined by personal experiences, perhaps as a result of previous studies. When many historical experiments similar to the experiment under study are available, these may guide the choice of a prior distribution from within a class, e.g., a specific gamma distribution. It is natural to use a conjugate prior distribution, that is to say a prior distribution for which the posterior distribution follows the same parametric form as the prior distribution (Gelman et al., 1995a). Conjugate families are often used in practice because of their mathematical convenience. In addition, conjugate prior distributions

can often be interpreted as additional data which provide a practical perspective (Gelman et al., 1995a).

The second approach to selecting the prior distribution is the objective approach. Investigators are often reluctant to provide information about the model parameters for fear of biasing or misleading the analysis. In such cases, one often applies so called "noninformative" prior distributions (Berger, 1985). A noninformative prior distribution can be set up either as a flat distribution or as a distribution with a big variance. A prior distribution is called an improper prior distribution, if its integral is infinite. In other words, we call a prior distribution proper if it does not depend on the data and integrates (or can be made to integrate) to 1 (Gelman et al., 1995a). Care must be taken with improper prior distribution to ensure that the selected prior distribution yields a proper posterior distribution (one with finite integral). Inferences may be sensitive to the choice of prior distribution when data provide little replication at the level of variation corresponding to a particular variance parameter. When data provide enough replication for precise estimation of some parameters, such as population parameters, the choice of prior distribution is not a critical issue since the likelihood dominates the prior distribution.

3.3.2 Posterior inference

3.3.2.1 Introduction

With the Bayesian approach, inferences for the model parameter $\boldsymbol{\theta}$ are obtained from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$. As indicated in the previous subsection, Bayesian analysis with large samples is similar to likelihood-based methods. The posterior distribution is approximately normal in large samples regardless of the prior distribution used. Thus, when the sample size is large, a posterior distribution can be characterized by summary statistics like the posterior mean and variance (Gelman et al., 1995a).

In finite samples, it is often difficult to determine $p(\boldsymbol{\theta} \mid \mathbf{y})$ analytically, especially if $\boldsymbol{\theta}$ is high dimensional. Numerical methods are often used to study $p(\boldsymbol{\theta} \mid \mathbf{y})$. For example, one can use numerical integration to compute the posterior mean $E(\boldsymbol{\theta} \mid \mathbf{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}$. It is increasingly common to rely on simulation methods, especially MCMC, to study the posterior distribution (see, for example, Chib and Greenberg, 1995). These methods approximate the posterior distribution by a collection of simulations from the posterior distribution. This approach provides great flexibility to address a range of inferential questions.

One focus of this thesis is the efficiency of various MCMC algorithms. Consequently MCMC plays a large role in the remainder of this chapter.

3.3.2.2 Markov chain Monte Carlo methods

The exploration of the joint posterior distribution is commonly carried out using Markov chain Monte Carlo (MCMC) methods. A Markov chain is a sequence of random variables generated from a transition distribution $q(\ldots)$ constructed such that the distribution of the next random variable to be generated depends only on the current state of the chain (Chib and Greenberg, 1995; Norris, 1997). The goal in a Bayesian analysis is to construct a Markov chain that converges to a stationary distribution that is the target posterior distribution. Then the Markov chain is simulated for a time until the point where the simulated draws resemble draws from the target distribution. Several criteria have been proposed to diagnose convergence (see, for example, Brooks and Gelman, 1998).

The transition distribution $q(x, y)$ is the conditional distribution of moving to point y starting from point x. It has been shown that a chain has a unique station distribution if the transition distribution must be constructed to satisfy irreducibility and aperiodicity conditions (Tierney, 1996; Norris, 1997). These conditions are explained below.

- irreducibility: the Markov chain can reach any non-empty set with positive probability in some number of draws from all starting points.

- aperiodicity: the movement of the Markov chain is not subject to regular periodic transitions (periodic chains can only reach certain restricted points of the sample space in any given step of the Markov chain simulation.).

Often Markov chains are constructed by specifying an initial distribution and a transition distribution. MCMC methods turn Markov chain theory around by trying to find a transition distribution such that the Markov chain converges to a known target distribution $\pi(.)$. Metropolis et al. (1953) and Hastings (1970) provide general approaches for constructing these chains. In the remainder of this section, we consider several issues associated with MCMC and the diagnosis of convergence.

3.3.2.3 The Gibbs sampler

The Gibbs sampler (Geman and Geman, 1984; Casella and George, 1992) is one of the best known of the MCMC methods. It is applied by decomposing the parameter vector $\boldsymbol{\theta}$ into (possibly univariate) subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 \ \ldots \ \boldsymbol{\theta}_p)$. Each iteration of the Gibbs sampler cycles through the full conditional posterior distribution of each subvector of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_i$, conditional on the values of all other subvectors, denoted as $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1} \ldots, \boldsymbol{\theta}_p)$. Thus, each subvector $\boldsymbol{\theta}_i$ is updated conditional on the latest value of the other components. It is often possible through the use of conjugate prior distributions to have every full conditional posterior distribution be a known parametric distribution, from which posterior samples can easily be drawn.

If we define $\boldsymbol{\theta}^{t-1}$ as the value of $\boldsymbol{\theta}$ at iteration $t-1$, then the Gibbs sampler makes the transition to $\boldsymbol{\theta}^t$ via the following product of conditional distributions.

$$p(\boldsymbol{\theta}_1^t \mid \boldsymbol{\theta}_{-1}^{t-1}, \mathbf{y})\ p(\boldsymbol{\theta}_2^t \mid \boldsymbol{\theta}_1^t, \boldsymbol{\theta}_{-(1,2)}^{t-1}, \mathbf{y})\ p(\boldsymbol{\theta}_3^t \mid \boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t, \boldsymbol{\theta}_{-(1,2,3)}^{t-1}, \mathbf{y})\ \ldots\ldots\ p(\boldsymbol{\theta}_p^t \mid \boldsymbol{\theta}_{-p}^t, \mathbf{y}),$$

where those components within () of $\boldsymbol{\theta}_{-(.)}$ are excluded from $\boldsymbol{\theta}$. The Gibbs sampler travels through low dimensional spaces (the full conditional distributions) to generate transitions in the higher dimensional space (the joint posterior distribution) by the *product of kernels principle* (Chib and Greenberg, 1995). As a result, the Gibbs sampler plays a significant role in practice due to its computational simplicity.

3.3.3 Metropolis-Hastings algorithms

The Metropolis-Hastings (M-H) algorithm (Metropolis et al., 1953; Hastings, 1970) is used when it is not possible to simulate from a desired distribution. As such it can be applied directly to the entire posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$, or to one or more intractable conditional distributions within a Gibbs sampler. We describe it in the latter context.

The M-H algorithm requires the specification of a jumping distribution, denoted by $J(y \mid x)$. To make the discussion concrete, suppose a full conditional posterior distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}, \mathbf{y})$ is the target distribution. The M-H algorithm chooses a jumping distribution to generate a candidate value $\boldsymbol{\theta}_i^*$, and then calculates a ratio of importance ratios

$$\alpha = \frac{p(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_{-i}, \mathbf{y})/J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i)}{p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}^*, \mathbf{y})/J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^*)} \tag{3.7}$$

to determine whether to accept or reject the candidate point $\boldsymbol{\theta}_i^*$. The candidate is accepted with probability $p = min(\alpha, 1)$, and then $\boldsymbol{\theta}_i^t = \boldsymbol{\theta}_i^*$. The M-H algorithm can be summarized as follows, assuming that a Markov chain is simulated for subvector $\boldsymbol{\theta}_i$ at the t iteration with the current value $\boldsymbol{\theta}_i^{t-1}$.

1. Generate $\boldsymbol{\theta}_i^*$ from $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^{t-1})$ and u from a uniform distribution $U(0,1)$

2. Set

$$\boldsymbol{\theta}_i^t = \begin{cases} \boldsymbol{\theta}_i^* & \text{when} \quad \alpha \geq u \\ \boldsymbol{\theta}_i^{t-1} & \text{when} \quad \alpha < u. \end{cases}$$

It is obvious that the ideal jumping distribution for the M-H algorithm is the target distribution itself, i.e., $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^{t-1}) = p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}, \mathbf{y})$ independent of the value $\boldsymbol{\theta}_i^{t-1}$, since then the ratio of importance ratios is exactly 1 and the candidates are always accepted. The Gibbs sampler is a special case of the M-H algorithms, where the jumping distributions are equal to the full conditional posterior distributions (the ideal jumping distribution). Note that when $J(. \mid .)$ is symmetric (i.e., $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^*) = J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i)$), it leads to a simpler ratio of importance ratios (and this is the original Metropolis implementation).

A common jumping distribution for the M-H algorithm is a normal distribution. For example, a random walk jumping distribution (Roberts, 1996) uses $J(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{t-1}) = N(\boldsymbol{\theta}^{t-1}, c^2\mathbf{V})$, where $\mathbf{V}$ is a variance matrix and the weight c may be used to optimize the algorithm's performance. Note that this normal jumping distribution is symmetric as long as $\mathbf{V}$ does not depend on $\boldsymbol{\theta}^{t-1}$, so the jumping distribution will cancel out in forming the ratio α. If the value of $\mathbf{V}$ depends on the current state, the normal jumping distribution is no longer symmetric in its arguments and the ratio of the jumping distributions at $\boldsymbol{\theta}_i^{t-1}$ and $\boldsymbol{\theta}_i^*$ plays a role in α.

It has been suggested that an optimal acceptance rate for the random walk jumping distribution is in the range 0.23 – 0.45 (Gelman et al., 1995b). Sometimes a pilot run is used to determine the variance of the jumping distribution so that the acceptance rate is in that range. For more details about the M-H algorithm, the choice of jumping distribution, and acceptance rates, see Gelman et al. (1995a), Gelman et al. (1995b), and Bennett et al. (1996).

3.3.4 Convergence criteria for MCMC

Assuming that the Markov chain has a stationary distribution, it will eventually converge to that distribution. One difficulty is determining at what point the simulations from the chain can be taken as representatives of the target distribution. The earlier,

transitory behavior is sometime known as "burn-in". The approach to diagnosing convergence that we use is based on multiple independent parallel Markov chains, which are initiated with overdispersed starting points (Gelman and Rubin, 1992). When convergence is reached, the values are evenly mixed in a narrow region (for example, see the plot for the parameter alpha.f in the top half of Figure 3.1). In order to eliminate the effect of the starting distribution, the first halves of sequences are discarded as burn-in and inferences for any posterior quantities are drawn based on the second halves of the sequences.

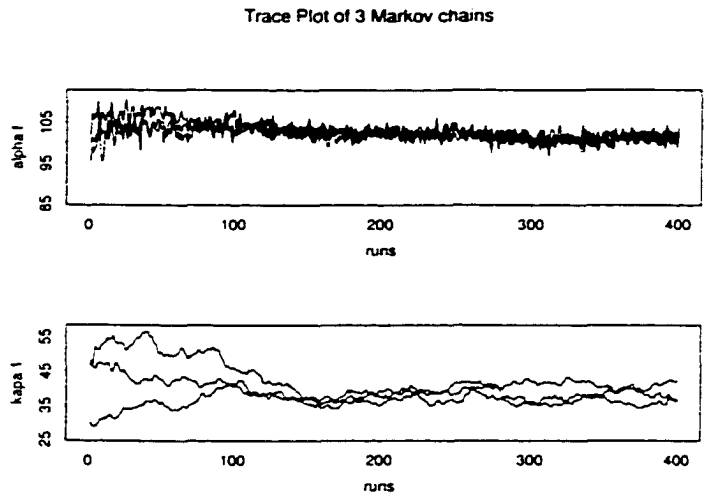

Figure 3.1 Time series plots of three Markov chains with different starting points. Top and bottom panel are for different parameters.

Convergence of multiple independent Markov chains can be checked by graphical, or numerical approaches (Brooks and Gelman, 1998). Plotting the time series of the posterior simulations of each parameter (as in Figure 3.1) to look for stationary behavior is a graphical way to check convergence. Plotting the posterior variance and within-sequence variance against iterations is another useful diagnostic. A number of

numerical measures are possible including exploration of autocorrelations between simulations, posterior correlation between parameters, or diagnostic measures based on means, variances, quantiles, or posterior interval lengths (Brooks and Gelman, 1998). Two numerical measures used for multiple Markov chains in this thesis are described next.

3.3.4.1 Potential scale reduction

Potential scale reduction (PSR, Gelman and Rubin, 1992) is derived from am adaptation of statistical analysis of variance. The basic idea is to compare the variability of simulations within chains (almost certainly an underestimate of posterior uncertainty) to a pooled posterior variance estimate (likely to be an overestimate). The PSR is a univariate measure estimated from the last halves of m independent chains, it is a variance ratio of pooled posterior variance to within-sequence variance,

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}} = \sqrt{\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}},$$

where n is the number of draws in the last half of each sequence, B is an estimate of between-sequence variance, W is an estimate of within-sequence variance, and the estimate of the pooled posterior variance is

$$\hat{V} = \frac{n-1}{n}W + \frac{m+1}{mn}B.$$

Large values for $\hat{R}$ suggest that the separate chains exhibit more variation than would be expected if each had converged to the target distribution. It can be shown that $\sqrt{\hat{R}}$ will approach 1.0 at convergence, with the pooled posterior variance close to the within-sequence variance. In most cases, a value of $\sqrt{\hat{R}}$ below 1.2 is acceptable (Gelman and Rubin, 1992), as this indicates that posterior inferences will become no more than 20% more precise if we continue the simulation. Sometimes the $\sqrt{\hat{R}}$ value may vary around 1.2 due to sampling variability before it remains constantly below 1.2. A $\sqrt{\hat{R}}$ plot against the iteration count can indicate when the sequences are well-mixed and converged.

3.3.4.2 Multivariate potential scale reduction

In many applications, multiple parameters are included in the models. Gelman and Rubin's PSR would have to be computed separately for each parameter and any other quantities of interest. As an alternative, Brooks and Gelman (1998) extended the definition of the PSR to the multi-dimensional case, yielding the multivariate potential scale reduction (MPSR), which combines the between-sequence covariance matrix $\mathbf{B}$ and within-sequence covariance matrix $\mathbf{W}$ for the parameters of interest into a scalar measure. MPSR should approach 1.0 as the convergence of multiple sequences is achieved. The estimated MPSR is denoted by $\sqrt{\hat{R}^p}$, and its definition and relation to the univariate PSR is given by

$$\begin{aligned}\sqrt{\hat{R}^p} &= \sqrt{max\ \frac{\mathbf{a}^T\hat{\mathbf{V}}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}} \\ &= \sqrt{\frac{n-1}{n}+\frac{m+1}{m}\lambda_1} \quad > max\sqrt{\hat{R}_i}\end{aligned}$$

where $\hat{\mathbf{V}} = \frac{n-1}{n}\mathbf{W} + \frac{m+1}{mn}\mathbf{B}$, λ_1 is the largest eigenvalue of $\mathbf{W^{-1}B}$, and $\sqrt{\hat{R}_i}$ are the univariate PSRs. Note that $\sqrt{\hat{R}^p}$ is an upper bound for the maximum $\sqrt{\hat{R}}$ of all univariate parameters. In other words, when $\sqrt{\hat{R}^p}$ <1.2, it indicates that all parameters of interest have converged. Thus when a symmetric positive definite matrix $\mathbf{W^{-1}B}$ exists, it is convenient to diagnose convergence using $\sqrt{\hat{R}^p}$.

The estimated MPSR, $\hat{R}^p$, is not calculable when either $\mathbf{B}$ or $\mathbf{W}$ is singular. If only $\mathbf{W}$ is singular, that suggests that at least one parameter has failed to move within the MCMC simulation. If both $\mathbf{B}$ and $\mathbf{W}$ are singular, then that suggests that at least two parameters are highly correlated. Examining plots of the determinants of $\hat{\mathbf{V}}$ and $\mathbf{W}$ helps to ensure that the chains are converging or to identify any underlying problems if they are not (Brooks and Gelman, 1998).

However, a limitation of the use of either PSR or MPSR is the assumption of (at least approximate) normality for the distribution of each parameter in the model, since the

diagnosis of convergence is based on means and variances alone. Without the assumption of normality, Brooks and Gelman (1998) introduce a family of potential scale reduction factors based on either posterior interval lengths or moments. Similar to the PSR based on analysis of variance, each of these PSR factors is interpreted as a measure of mixing of Markov chains, approaching 1 as convergence is achieved.

3.3.4.3 Measuring the convergence rate

In order to compare algorithms, we need a rule for declaiming that a particular MCMC has converged. Our process is to run m independent chains for a long time, and then evaluate $\sqrt{\hat{R}^p}$, the estimate of MPSR, every s iterations (e.g., $s = 1000$), using the last halves of the sequences. That is, we evaluate $\sqrt{\hat{R}^p}$ at iterations 1.000, 2.000, 3.000, ... , by using the last half of every sequence (i.e., the value for n in $\sqrt{\hat{R}^p}$ is 500, 1.000, 1.500...). The convergence point of an algorithm, denoted by γ (a multiple of s), is the first iteration for which $\sqrt{\hat{R}^p}$ goes below 1.2 and continuously stays below 1.2 for 20.000 iterations. We use γ as an indicator for the efficiency of the MCMC algorithm. The higher the convergence rate is, the smaller the value of γ will be. Thus, an algorithm is preferred to others if it yields the smallest γ. Note that the convergence point is not necessarily indicated on a fine scale, since γ must be an integral multiple of the choice of s.

3.4 Comments on the likelihood-based and Bayesian approaches

In the last section of this chapter we discuss the relationship of the likelihood and Bayesian approaches. Historically animal breeders have relied on REML-BLUP analysis to estimate variance components, fixed effects, and animal breeding values. Though computationally difficult for large data sets, it has been the main method in widespread use. More recently (Gianola and Fernando, 1986; Rodriguez-Zas et al., 1998; Blasco et

al., 1998; Wright et. al., 2000) Bayesian methods, with the advent of MCMC methods, have presented an appealing alternative for model fitting in animal breeding.

There is a close relationship between the Bayesian and REML-BLUP approaches. From a Bayesian point of view, the REML estimate of variance components $\boldsymbol{\omega}$ is the mode of the marginal posterior density of $\boldsymbol{\omega}$, which is proportional to the product of the likelihood function l_1 (integrating over random effects) and a uniform prior distribution for ω (Harville, 1977).

For the linear mixed model (3.2) with normally distributed random effects and errors, Robinson (1991) and Harville (1991) point out that (empirical) REML-BLUP estimates are equivalent to (parametric empirical) Bayesian estimates (posterior means). Specifically, assuming a joint normal distribution of $\mathbf{y}$ and $\mathbf{u}$, as well as a uniform improper prior distribution for $\mathbf{b}$, the mode of the joint posterior distribution of $\mathbf{b}$ and $\mathbf{u}$ given the variance components is given by the solution to Henderson's MME.

There are several nice features of the Bayesian approach to animal breeding problems. First, these inferences do not depend on asymptotic results. Second, it makes it easy to obtain flexible inferences for any quantity of interest (e.g., rankings, heritability) from the joint posterior distributions of parameters. For example, confidence intervals for heritability can be calculated from the joint posterior distribution of the corresponding variance components without relying on asymptotic properties and the delta method. Third, the posterior distribution of random effect parameters automatically accounts for uncertainty in estimating the variance component parameters. Fourth, the Bayesian approach is flexible enough to accommodate missing data or nonstandard likelihood functions. Finally, the Bayesian approach can make it easy to work with large data sets since it is possible to avoid inversion of large matrices (i.e., $var(\mathbf{y}) = \mathbf{V}$).

There are also disadvantages to applying Bayesian approach. First and foremost is the need for specification of prior distributions for model parameters. Investigators often elect not to use subjective information. There is an attempt to use noninformative prior

distributions, though these can be somewhat arbitrary and may lead to improper posterior distributions without any obvious signs of failure of MCMC. A second disadvantage is the difficulty of detecting convergence of MCMC algorithms.

CHAPTER 4 SOME ISSUES IN IMPLEMENTING BAYESIAN METHODS FOR LINEAR AND NONLINEAR MODELS

4.1 Introduction

The discussion of Bayesian methods in Chapter 3 outlines the three steps required for a Bayesian analysis: First, set up a probability model, including the probability distribution for the response conditional on parameters and prior distributions for the model parameters. Second, calculate the joint posterior distribution of the quantities of interest. Finally, evaluate the fit of the model. There are many choices to be made in each of these steps. In this chapter, several issues related to carrying out the first two steps are discussed.

The layout for this chapter is as follows. The remainder of this section introduces the issues that affect the efficiency of Bayesian methods. Two strategies used to improve the efficiency of Bayesian methods for random regression models are introduced in Sections 4.2 and 4.3. In Section 4.4, different Metropolis-Hastings algorithms are proposed for analyzing nonlinear models. Some other issues associated with improving efficiency of Bayesian methods are briefly discussed in the final section of this chapter.

4.1.1 Efficiency of Bayesian methods

Bayesian inferences are based on the joint posterior distribution of the model parameters, which is proportional to the product of the prior distributions and the likelihood function (see Section 3.3). In a Bayesian analysis, the exploration of the joint posterior distribution often depends on the application of Markov chain Monte Carlo (MCMC) methods (see Section 3.3.2.2). These methods generate Markov chains from a transition distribution constructed such that the Markov chain converges to the target posterior distribution (Chib and Greenberg, 1995; Tierney, 1996; Norris, 1997). When the MCMC algorithm has converged, its simulations are representative of the target distribution and not influenced by the starting points for the Markov chains.

The number of iterations required for the Markov chains to have converged can be used as a measure for the efficiency of the MCMC method. The smaller the number of simulations required to obtain convergence, the more efficient the MCMC method. Other factors are surely relevant, for example, the ease of programming, the amount of computer time required for each iteration, and the number of approximately independent samples in a fixed length simulation. In this chapter, the number of iterations required for convergence is our main concern.

4.1.2 Some methods for assessing convergence rate

In our study, convergence rate is measured by recording the total number of iterations required for the Markov chain to have converged (see Section 3.3.4.3). The higher the convergence rate is, the smaller the number of iterations required it to converge. Measuring the mixing of Markov chains is an informal way to look at the convergence rate. The relative magnitude of between- and within-sequence variance is often used to detect mixing when multiple Markov chains with the same stationary with the same stationary distribution are simulated (see Section 3.3.4). Autocorrelations of the posterior

simulations can also be used for assessing mixing. If autocorrelation is high, posterior draws will stay in a narrow region of the parameter space for a long time before moving to another region. Therefore, the Markov chain needs a large number of iteration to travel the entire parameter space. Autocorrelation can be reduced by transforming parameters.

Examining the correlation among parameters, calculated from posterior samples, is another way to detect slow mixing of MCMC (Roberts, 1996; Gilks and Roberts, 1996). A reduction in the posterior correlation can enlarge the step between successive draws and hence improve mixing. For example, consider a simple linear regression model, $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Assume that σ^2 is known and flat prior distributions are assumed for parameters β_0 and β_1. Then the posterior correlation between β_0 and β_1 is

$$\rho(\beta_0, \beta_1) = \frac{-\bar{x}}{\sqrt{\bar{x}^2 + \sum(x_i - \bar{x})^2/n}}.$$

We can observe that β_0 and β_1 are uncorrelated if $\bar{x} = \sum_i^n x_i/n = 0$. This model can be reparameterized by centering the covariate, i.e., define $x^* = x - \bar{x}$. This results in $\bar{x}^* = 0$ and independence between the new parameters $\beta_0^* = \beta_0 + \beta_1 \bar{x}$ and $\beta_1^* = \beta_1$. The independence is implied by the zero covariance because the joint posterior distribution for β_0^* and β_1^* is normal. Hence,using this reparameterized model, the Gibbs sampler, which works on full conditional distributions, will immediately produce samples from the posterior distribution without any burn-in.

4.1.3 Factors affecting the convergence rate

Many factors can be manipulated in the model specification step or in developing a computing strategy to improve the MCMC convergence rate. After identifying the factors related to convergence here, we discuss methods for improving the convergence rate in more detail in the remaining sections of this chapter. The previous section showed

the impact of autocorrelation and posterior correlation on the efficiency of MCMC algorithms. Reparameterization is a key tool in developing efficient algorithms.

One model is considered a reparameterization of another if the parameters of the model in question may be expressed as a function only of the parameters of the other model, without the expression containing the explanatory variables, the response variables, or the error term (Ratkowsky, 1990). For example, the additive model for a two-way crossed classification with main effects and interactions, $y_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ is a reparameterization of the model $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$, where $\mu_{ij} = \alpha_i + \beta_j + \alpha\beta_{ij}$. The different parameterizations, $(\alpha_i\ \beta_j\ \alpha\beta_{ij})$ versus (μ_{ij}) results in different conditional posterior distributions and consequently different MCMC algorithms (Gelfand et al., 1995). Since MCMC methods are often applied to the conditional posterior distributions (e.g., the Gibbs sampler), an increase in mixing occurs when models are parameterized in terms of independent components, since then the conditional distributions do not depend on some of the variables being conditioned on. Although different parameterizations of the same model produce the same fitted values, they may differ greatly in the efficiency of the MCMC algorithms that is produced (Gelfand and Carlin, 1995; Gelfand et al., 1995).

When the Metropolis-Hastings algorithm is required (e.g. for nonlinear models), the choice of jumping distribution can have a large effect on convergence rate. The format of jumping distribution (Gilks, 1996), the size of the typical move (Gelman et al., 1995b), and the relationship of the jumping distribution to the posterior distribution can affect the convergence rate.

In addition, the starting points used in an MCMC analysis can affect the convergence rate, especially for slow-mixing chain (Gilks et al., 1996). Whether model parameters are drawn element-by element or in batches is another factor that can affect the convergence rate. Batching parameters into a single joint distribution during MCMC implementation allows for movement about the parameter space more quickly than a series of single-

element steps, especially when some parameters are highly correlated (Gilk and Roberts, 1996). In the remainder of this chapter, we expand on the ideas described in this first section.

4.2 Hierarchical centering

4.2.1 Introduction

Hierarchical models, or random effects models, including random regression models for analyzing animal growth data, describe response variables in terms of individual parameters that are assumed to vary around population mean values. There is more than one way to parameterize such models. We demonstrate by considering a simple linear mixed model, $y_{ij} = \mu + u_i + \epsilon_{ij}$, with independent, identically distributed (iid) error observations, $\epsilon_{ij} \sim N(0, \sigma_e^2)$, and random effects $u_i \sim N(0, \sigma_u^2)$. Note that the random effects are assumed randomly distributed with mean 0 in this parameterization. A reparameterized version of the model is $y_{ij} = u_i^* + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma_e^2)$, and $u_i^* = \mu + u_i \sim N(\mu, \sigma_u^2)$. The two models are clearly equivalent. In the second version, the u_i^*'s are said to be centered parameters and this reparameterization is known as centering. The u_i's in the initial model are called uncentered parameters. Gelfand and Carlin (1995) and Gilks and Roberts (1996) show that centering of this type can reduce correlations between low dimensional parameter subvectors and hence speed up convergence.

4.2.2 Hierarchical centering for basic models

Guidelines for determining whether hierarchical centering will improve convergence in the simple mixed model with two variance components are summarized by Gelfand et al. (1995a) as follows. Let $\mathbf{G}$ denote the variance of the uncentered parameters (i.e., $u_i \sim N(0, \mathbf{G})$) and $\mathbf{B}$ the variance of the conditional posterior distribution of the centered

parameters (i.e., $\boldsymbol{\eta}_i = \boldsymbol{\mu}+\mathbf{u}_i$). If $|\mathbf{BG}^{-1}|$ is near zero, then the centering parameterization is efficient, while if $|\mathbf{BG}^{-1}|$ is near one then the uncentered parameterization will be preferred. To illustrate, Gelfand et al. (1995a) consider the simple balanced mixed model,

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \alpha_i \sim N(0, \sigma_\alpha^2)$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $i = 1, \ldots, n$; $j = 1, \ldots, r$. Centering the random effects α_i at μ rather than at zero gives

$$y_{ij} = \eta_i + \epsilon_{ij}, \qquad \eta_i = \mu + \alpha_i \sim N(\mu, \sigma_\alpha^2),$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The variance of the conditional posterior distribution of the centered parameter η_i conditional on $\mathbf{y}_i, \mu, \sigma_\alpha^2, \sigma_\epsilon^2$ is $(1/\sigma_\alpha^2 + 1/\sigma_\epsilon^2)^{-1} = B$. We observe that the ratio of the variance of the conditional posterior distribution for η_i to the variance of α_i is $|\mathbf{BG}^{-1}| = B/\sigma_\alpha^2 = \sigma_\epsilon^2/(\sigma_\epsilon^2 + \sigma_\alpha^2)$, which approaches zero when σ_α^2 is large relative to σ_ϵ^2. In that case, centering is effective and η_i quickly approaches its correct distribution. Thus, for the simple balanced mixed model, when the variability at higher levels of the hierarchical model (σ_α^2) dominates that at the lower levels (σ_ϵ^2), hierarchical centering appears to be helpful.

Gilks and Roberts (1996) also use this simple balanced mixed model to provide further insight into the situation where centering can improve convergence. They show that the posterior correlations between uncentered random effects are

$$\rho_{\mu,\alpha_i} = -\sqrt{\frac{\sigma_\alpha^2/n}{\sigma_\alpha^2/n + \sigma_\epsilon^2/r}} \quad \text{and} \quad \rho_{\alpha_i,\alpha_k} = \frac{\sigma_\alpha^2/n}{\sigma_\alpha^2/n + \sigma_\epsilon^2/r} \qquad \text{for } i \neq k.$$

The correlations among the α_i's are large when $\sigma_\alpha^2/n \gg \sigma_\epsilon^2/r$, i.e., when σ_α^2 is large. The posterior correlations between the centered random effects η_i are

$$\rho_{\mu,\eta_i} = \sqrt{\frac{\sigma_\epsilon^2}{n r \sigma_\alpha^2 + \sigma_\epsilon^2}} \quad \text{and} \quad \rho_{\eta_i,\eta_k} = \frac{\sigma_\epsilon^2}{n r \sigma_\alpha^2 + \sigma_\epsilon^2} \qquad \text{for } i \neq k.$$

When σ_α^2 is large, these correlations are small, resulting in improved convergence. These results support those of Gelfand et al. (1995a). There is no advantage to centering

random effects when their variance is small. When $\rho_{\eta_i,\eta_k} \ll \rho_{\alpha_i,\alpha_k}$: that is, $\sigma_e^2 \ll r\sigma_\alpha$, then the mixing rate will be faster in the centered model. Hence, if σ_α^2 is big, centering the random effects will improve convergence. For the simple mixed model, the best parameterization depends on the magnitude of the variance of the random effects relative to the variance of the residuals.

4.2.3 Hierarchical centering for linear mixed models

The mixed models applied in animal breeding are more sophisticated than the simple model described in the previous section. The idea described above can be easily generalized as long as there are only two vectors of random effects per individual unit. For example, consider the linear mixed model (2.3) in Chapter 2. Here we write the model in terms of the n_i observations for animal i.

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{y}_i$ is the vector of observations for individual i, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are incidence matrices associated with fixed effects $\mathbf{b}$ and random effects $\mathbf{u}_i$, respectively, $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G})$, and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{I}_{n_i}\sigma_e^2)$. If $\mathbf{X}_i = \mathbf{Z}_i$ (as occurs with our random regression model in the same order polynomial is used in both parts of the model), then the hierarchical centered model is

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i.$$

where $\boldsymbol{\eta}_i = \mathbf{b} + \mathbf{u}_i \sim N(\mathbf{b}, \mathbf{G})$. Assuming a flat prior distribution for $\mathbf{b}$, the posterior variance for the centered parameters $\boldsymbol{\eta}_i$ is $\mathbf{B} = (\mathbf{Z}_i^T\mathbf{Z}_i/\sigma_e^2 + \mathbf{G}^{-1})^{-1}$. The principle of Gelfand et al. (1995) can be easily applied by looking at the ratio of the posterior variance of the centered parameters to the variance of the uncentered parameters. The ratio is $|\mathbf{B}\mathbf{G}^{-1}| = \left|\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{G}/\sigma_e^2 + \mathbf{I}\right|^{-1}$. If $\left|\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{G}/\sigma_e^2 + \mathbf{I}\right|$ is large, hierarchical centering would be preferred over the uncentered linear mixed model. In other words, if σ_e^2 is small and $|\mathbf{G}|$ is large, then centering the random effects will improve convergence in

the MCMC algorithm. Interestingly, the form of the posterior variance matrix $\mathbf{B}$ also suggests possible benefits to orthogonal $\mathbf{Z}$. We will discuss this in the next section.

When a model includes more than two variance components, as in the animal model with genetic effects, permanent environmental effects, and errors (model (2.1)), the arguments above do not apply directly. Gelfand et al. (1995a) recommend that the effects of having the largest posterior variance relative to the variance of the residuals σ_e^2 should be centered. The authors further suggest estimating the posterior expectation of the variance components from a preliminary MCMC run in order to compare the relative magnitudes of variance components. The heritability for most growth traits is less than 0.5; for example, the heritability for pig weight at 150 to 180 days is in the range 0.2 to 0.35. Therefore, it seems likely that centering the permanent environmental effects rather than the animal genetic effects will be most effective in the applications considered later. We compare the approaches using simulation in Chapter 5.

In some cases it may be difficult to determine which parameterization will be most efficient for a given model. Gelfand and Carlin (1995) introduced the cycling MCMC algorithm to address such situations. The cycling algorithm consists of a cycling through all of the relevant parameterizations in sequence; that is, separate MCMC algorithm are developed for each parameterization, and then one complete iteration (all parameters) is run for each parameterization in sequence. This can achieve the benefit of hierarchical centering, without requiring the user to select the correct parameterization.

Note that centering can also be applied to nonlinear models, when the parameters of the nonlinear functions have hierarchically linear structure (e.g., $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \delta_i \sim \mathbf{N}(\boldsymbol{\theta}_0, \mathbf{G})$). The principles outlined in this section can provide guidance for such models. But as the models are nonlinear, the arguments of Gelfand et al. (1995) do not directly apply.

4.3 Orthogonal polynomials

We will now discuss a transformation, the use of orthogonal polynomials in place of the ordinary, that can be used to improve the efficiency of MCMC in random polynomial regression models.

4.3.1 Definition

Two vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be orthogonal vectors when $\mathbf{u}^T\mathbf{v}=0$. The vectors are called orthonormal if in addition $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v}=1$. A matrix $\mathbf{A}$, whose columns constitute a set of orthonormal vectors, is called an orthogonal matrix (Searle, 1971); it satisfies $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. These notions can be extended to polynomials and the resulting orthogonal polynomials are useful for our random regression model.

Let $P_i(t)$ be a polynomial with non-zero coefficient for term t^i and $F(t)$ be a non-negative weight function. Then a system of polynomials $P_i(t)$ $(i = 0, 1, \ldots)$ are orthogonal on the interval (a, b) with respect to weight function $F(t)$ if $\int_a^b P_i(t)P_j(t)F(t)dt = 0$ (Thisted, 1988). It is common to standardize the polynomials such that $\int_a^b P_i(t)^2F(t)dt = 1$. Many families of orthogonal polynomials satisfy a recurrence relation of the form,

$$P_j = (a_j + b_jt)P_{j-1} - c_jP_{j-2}. \tag{4.1}$$

where a_j, b_j, and c_j identify the family (Thisted, 1988). An example we apply later are the Legendre polynomials, which are described in detail in Chapter 5.

4.3.2 Rationale for using orthogonal polynomials

The Gibbs sampler draws samples from the full conditional posterior distribution of each parameter (or subvector of parameters) in sequence. If some parameters in the model are highly correlated in the posterior distribution, then posterior draws from the full conditional posterior distribution of such parameters will tend to be from a narrow

range of the parameter space, which results in slow mixing of the Markov chains (Gilks and Roberts, 1996). That is a potential difficulty for the random regression models proposed here because coefficients from regular polynomial regression models are generally highly correlated. Transformation to orthogonal polynomials leads to coefficients that are uncorrelated with each other, and, therefore, the conditional distribution of the p^{th} polynomial coefficient is independent of the coefficients of the other polynomial coefficients. For example, consider a simple polynomial model, $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ with $\mathbf{X}$ corresponding to polynomial terms (e.g., $1, x, x^2, x^3, \ldots$), $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, and a noninformative prior distribution for $\mathbf{b}$. For such a model the posterior variance of $\mathbf{b}$ is: $var(\mathbf{b} \mid \mathbf{y}, \sigma^2) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. If orthogonal polynomials are used with $\mathbf{X}^*$, denoting the design matrix for the orthogonal polynomials, then $var(\mathbf{b} \mid \mathbf{y}, \sigma^2) = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\sigma^2 = \mathbf{I}\sigma^2$, and the components of $\mathbf{b}$ are independent in their posterior distribution.

For the animal model, with balanced repeated records on each animal, $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$ with $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. When fitting fixed effects $\mathbf{b}$ with p^{th} degree polynomial and random effects $\mathbf{u}$ with a q^{th} ($q \leq p$) degree polynomial, the hierarchical centered model is $\mathbf{y} = \mathbf{X}^+\mathbf{b}^+ + \mathbf{Z}\boldsymbol{\eta} + \mathbf{e}$ with $\mathbf{X} = (\mathbf{Z}\ \mathbf{X}^+)$, $\mathbf{b} = (\tilde{\mathbf{b}}\ \mathbf{b}^+)$, and $\boldsymbol{\eta} = \tilde{\mathbf{b}} + \mathbf{u}$ (see Section 5.3 for more details). Note $\mathbf{X}^+$ contains the higher order polynomial terms not included as random effects. Assume a $N(\tilde{\mathbf{b}}, \mathbf{G})$ prior distribution for $\boldsymbol{\eta}$ and a flat prior distribution for $\mathbf{b}^+$. With orthogonal design matrices $\mathbf{X}^+$ and $\mathbf{Z}$, the joint conditional (on σ_e^2 and $\mathbf{G}$) posterior distribution of $\mathbf{b}^+$ and $\boldsymbol{\eta}$ is.

$$
\begin{aligned}
p(\mathbf{b}^+, \boldsymbol{\eta} \mid \mathbf{y}, \sigma_e^2) &\propto p(\mathbf{y} \mid \mathbf{b}^+, \boldsymbol{\eta}, \sigma_e^2)\, p(\mathbf{b}^+, \boldsymbol{\eta}) \\
&\propto exp\left(\frac{-1}{2\sigma_e^2}(\mathbf{y} - \mathbf{X}^+\mathbf{b}^+ - \mathbf{Z}\boldsymbol{\eta})^T(\mathbf{y} - \mathbf{X}^+\mathbf{b}^+ - \mathbf{Z}\boldsymbol{\eta})\right) \\
&\qquad exp\left(\frac{-1}{2}(\boldsymbol{\eta} - \tilde{\mathbf{b}})^T\mathbf{G}^{-1}(\boldsymbol{\eta} - \tilde{\mathbf{b}})\right) \\
&\quad \Downarrow \mathbf{X}^+ \text{ and } \mathbf{Z} \text{ are orthogonal, so } \mathbf{X}^{+T}\mathbf{Z} = \mathbf{0} \\
&\propto exp\left(\frac{-1}{2\sigma_e^2}(\mathbf{b}^{+T}\mathbf{X}^{+T}\mathbf{X}^+\mathbf{b}^+ - \mathbf{y}^T\mathbf{X}^+\mathbf{b}^+)\right)
\end{aligned}
$$

$$exp\left(\frac{-1}{2}\left((\boldsymbol{\eta}^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\eta} - \mathbf{y}^T\mathbf{Z}\boldsymbol{\eta})/\sigma_e^2 - (\boldsymbol{\eta} - \tilde{\mathbf{b}})^T\mathbf{G}^{-1}(\boldsymbol{\eta} - \tilde{\mathbf{b}})\right)\right)$$

$$\propto \mathcal{N}(\mathbf{b}^+ \mid \hat{\mathbf{b}}, \hat{\Sigma}_b)\, \mathcal{N}(\boldsymbol{\eta} \mid \hat{\boldsymbol{\eta}}, \hat{\Sigma}_\eta)$$

where $\hat{\mathbf{b}}^+ = (\mathbf{X}^{+T}\mathbf{X}^+)^{-1}\mathbf{X}^{+T}\mathbf{y} = \mathbf{X}^{+T}\mathbf{y}$.

$\hat{\Sigma}_b = (\mathbf{X}^{+T}\mathbf{X}^+)^{-1}\sigma_e^2 = \mathbf{I}\sigma_e^2$. $\hat{\boldsymbol{\eta}} = \hat{\Sigma}_\eta(\mathbf{Z}^T\mathbf{y}/\sigma_e^2 + \mathbf{G}^{-1}\tilde{\mathbf{b}})$.

and $\hat{\Sigma}_\eta = (\mathbf{Z}^T\mathbf{Z}/\sigma_e^2 + \mathbf{G}^{-1})^{-1} = (\mathbf{I}/\sigma_e^2 + \mathbf{G}^{-1})^{-1}$.

$$= p(\mathbf{b}^+ \mid \mathbf{y}, \sigma_e^2)\, p(\boldsymbol{\eta} \mid \mathbf{y}, \sigma_e^2, \mathbf{G})$$

$$= p(\mathbf{b}^+_{q+1} \mid \mathbf{y}, \sigma_e^2)\, p(\mathbf{b}^+_{q+2} \mid \mathbf{y}, \sigma_e^2) \ \ldots\ p(\mathbf{b}^+_p \mid \mathbf{y}, \sigma_e^2)\, p(\boldsymbol{\eta} \mid \mathbf{y}, \sigma_e^2, \mathbf{G})$$

Therefore, when $\mathbf{X}^+$ and $\mathbf{Z}$ consist of orthogonal vectors and $\mathbf{X}^+\mathbf{Z} = \mathbf{0}$, the posterior distributions for $\mathbf{b}^+$ and $\boldsymbol{\eta}$ are independent. Because the orthogonality eliminates the correlation between parameters in the conditional posterior distributions, the convergence rate of the Markov chain simulations is improved.

4.4 Metropolis-Hastings algorithms

The Metropolis-Hastings (M-H) algorithm (Metropolis et al., 1953; Hastings, 1970) is commonly used when it is not possible to sample directly from a distribution of interest (either the entire posterior distribution or a single conditional distribution within a Gibbs sampler). The M-H algorithm requires a jumping distribution to obtain candidate values, and then calculates the ratio of importance ratios α in (3.7) to determine whether the candidates are accepted or rejected (Chib and Greenberg, 1995). The M-H algorithm is required for the nonlinear models we apply in Chapter 6. This section reviews issues associated with the choice of jumping distribution that can affect the MCMC convergence rate.

4.4.1 Linearization of nonlinear models

The standard approach to working with nonlinear models, e.g., parameter estimation, relies on repeated linear approximation to the nonlinear function (Ratkowsky and Dolby, 1975; Sheiner and Beal, 1980; Lindstrom and Bates, 1990). Let $f(\boldsymbol{\theta})$ denote a nonlinear function with parameter vector $\boldsymbol{\theta}$. Expanding $f(\boldsymbol{\theta})$ in a first-order Taylor series around a value $\boldsymbol{\theta}^*$ yields a linear approximation

$$f(\boldsymbol{\theta}) \approx f(\boldsymbol{\theta}^*) + \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*).$$

Suppose that we work with the model

$$y_i = f(\boldsymbol{\theta}, t_i) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_e^2)$. The log-likelihood function of n observations is proportional to

$$\begin{aligned} l(\boldsymbol{\theta}) &= log\left(L(\boldsymbol{\theta} \mid \mathbf{y})\right) = log\left(\Pi_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}, t_i)\right) = \sum_{i=1}^{n} log\left(N(y_i \mid f(\boldsymbol{\theta}, t_i), \sigma_e^2)\right) \\ &\propto exp\left(\frac{-1}{2\sigma_e^2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{\theta}, t_i))^2\right). \end{aligned}$$

Note that this log-likelihood function is not quadratic in $\boldsymbol{\theta}$. Applying the Taylor series expansion of the nonlinear function in the log-likelihood function yields a quadratic approximation (in $\boldsymbol{\theta}$) to the log-likelihood function.

$$l(\boldsymbol{\theta}) \approx exp\left\{\frac{-1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i - f(\boldsymbol{\theta}^*, t_i) - \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right)^2\right\}.$$

The approximate likelihood is of the form of a multivariate normal distribution with respect to $\boldsymbol{\theta}$. Let $\mathbf{F} = \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$, then

$$\begin{aligned} l(\boldsymbol{\theta}) &\approx exp\left\{\frac{-1}{2\sigma_e^2}\sum_{i=1}^{n}\left[\mathbf{F}^T\boldsymbol{\theta} - \left(y_i - f(\boldsymbol{\theta}^*, t_i) + \mathbf{F}^T\boldsymbol{\theta}^*\right)\right]^2\right\} \\ &\propto exp\left\{\frac{-1}{2\sigma_e^2}\sum_{i=1}^{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T\mathbf{F}\mathbf{F}^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\} \\ &\propto N\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, \hat{\Sigma}_{\boldsymbol{\theta}}\right), \end{aligned} \tag{4.2}$$

where $\hat{\Sigma}_{\boldsymbol{\theta}} = (\mathbf{F}\mathbf{F}^T)^{-1}\sigma_\epsilon^2$ and $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \hat{\Sigma}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathbf{F}\,(y_i - f(\boldsymbol{\theta}^*, t_i))\,/\sigma_\epsilon^2$

In the M-H context, one can linearize around the current value $\boldsymbol{\theta}_i^{\wedge}$. This yields a normal jumping distribution that is expected to be close to the target distribution (Bennett et al., 1996).

4.4.2 Choice of jumping distribution

Any jumping distribution that satisfies the MCMC regularity conditions (irreducibility and aperiodicity) converges to a unique stationary distribution (see Section 3.3.2.2). We would like to choose a jumping distribution to be close to the posterior distribution under study. For example, the full conditional posterior distribution itself is used as the jumping distribution for the Gibbs sampler algorithm.

Gelman et al. (1995b) discuss efficient Metropolis jumping rules. Bennett et al. (1996) have reviewed several M-H algorithms used for nonlinear hierarchical models. Here we will review a number of possible choices for jumping distributions. Several of these are applied to an example in Chapter 6.

Before introducing algorithms, we define the following notation.

Notation

$\boldsymbol{\theta}_i^{\wedge}$: the current value for $\boldsymbol{\theta}_i$ at a certain iteration of the Markov chain;

$\boldsymbol{\theta}_i^*$: a candidate value for $\boldsymbol{\theta}_i$ drawn from the jumping distribution;

$\boldsymbol{\theta}_i^{MLE}$: the maximum likelihood estimate for $\boldsymbol{\theta}_i$ obtained by fitting the nonlinear model only to data from individual i;

$p(\boldsymbol{\theta}_i \mid \mathbf{y})$: the posterior distribution for $\boldsymbol{\theta}_i$;

$p(\boldsymbol{\theta}_i)$: the prior distribution for $\boldsymbol{\theta}_i$;

$p(\mathbf{y} \mid \boldsymbol{\theta}_i)$: the sampling distribution of $\mathbf{y}$;

$J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c)$: a jumping distribution with argument $\boldsymbol{\theta}_i$ conditional on the value $\boldsymbol{\theta}_i^c$:

α: the ratio of importance ratios

$$\alpha = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^*)\, p(\boldsymbol{\theta}_i^*)/J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i^c)}{p(\mathbf{y} \mid \boldsymbol{\theta}_i^c)\, p(\boldsymbol{\theta}_i^c)/J(\boldsymbol{\theta}_i^c \mid \boldsymbol{\theta}_i^*)} = \frac{p(\boldsymbol{\theta}_i^* \mid \mathbf{y})/J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i^c)}{p(\boldsymbol{\theta}_i^c \mid \mathbf{y})/J(\boldsymbol{\theta}_i^c \mid \boldsymbol{\theta}_i^*)}. \tag{4.3}$$

Suppose the goal is to sample from $p(\boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{\theta}_{-i})$ as part of a Gibbs sampling algorithm, where $\boldsymbol{\theta}_{-i}$ contains all parameters except $\boldsymbol{\theta}_i$. Recall the M-H algorithm operates as follows: a candidate value $\boldsymbol{\theta}_i^*$ is generated from $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c)$. The candidate $\boldsymbol{\theta}_i^*$ is accepted with probability equal to $\min(\alpha, 1)$. If the candidate is rejected, we carry forward $\boldsymbol{\theta}_i^c$.

We now introduce four M-H algorithms. Note that the differences between these algorithms include: (i)whether the jumping distribution depends on the forms of the posterior distribution, and (ii)whether the mean or variance of the jumping distribution depends on the current state.

Algorithm I: Independent M-H algorithm

Perhaps the simplest M-H algorithm is the independence M-H algorithm, in which the jumping distribution does not depend on the current state. In the independent random walk M-H algorithm, candidates are drawn from a distribution which is independent of the current value $\boldsymbol{\theta}_i^c$. This approach is introduced by Smith and Gelfand (1992) and Tierney (1994) and is motivated by the case in which data are sparse and it seems reasonable to ignore it in generating candidates. The choice of a constant jumping distribution leads to considerable simplification of the ratio of importance ratios. For example, using the prior as a jumping distribution, the ratio of importance ratios is

$$\alpha = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^*)\, p(\boldsymbol{\theta}_i^*)/p(\boldsymbol{\theta}_i^*)}{p(\mathbf{y} \mid \boldsymbol{\theta}_i^c)\, p(\boldsymbol{\theta}_i^c)/p(\boldsymbol{\theta}_i^c)} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_i^*)}{p(\mathbf{y} \mid \boldsymbol{\theta}_i^c)}.$$

which only depends on the ratio of the sampling distributions. This algorithm will not be used in our subsequent work.

Perhaps the most popular M-H algorithm is the random walk M-H algorithm (Metropolis, 1953; Chib and Greenberg, 1995), because the candidate is equal to the current value plus a disturbance. There are two types described below: Algorithm II and Algorithm III.

Algorithm II: Random walk M-H algorithm

By taking the jumping distribution to be normal with mean equal to the current value and variance proportional $\boldsymbol{\Sigma}$, $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c) = N(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c, c\boldsymbol{\Sigma})$, we have a random walk in $\boldsymbol{\theta}_i$ space. The key assumption is that $\boldsymbol{\Sigma}$ does not depend on $\boldsymbol{\theta}_i^c$. The constant c is included because it is sometimes desirable to adjust the variance to improve the convergence rate. If one uses a random walk jumping distribution of this form which is symmetric in $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i^c$, then the ratio α is of a simple form:

$$\alpha = \frac{p(\boldsymbol{\theta}_i^* \mid \mathbf{y})/J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i^c)}{p(\boldsymbol{\theta}_i^c \mid \mathbf{y})/J(\boldsymbol{\theta}_i^c \mid \boldsymbol{\theta}_i^*)} = \frac{p(\boldsymbol{\theta}_i^* \mid \mathbf{y})}{p(\boldsymbol{\theta}_i^c \mid \mathbf{y})}.$$

It is observed that the size of the step from $\boldsymbol{\theta}_i^c$ to $\boldsymbol{\theta}_i^*$ does not depend on the current point $\boldsymbol{\theta}_i^c$ at all. Note that α depends only on the posterior distributions if the jumping distribution is symmetric in its arguments.

Algorithm III: Scale-dependent random walk M-H algorithm

Depending on the form of the posterior distribution, we may want to let the scale matrix depend on the current value $\boldsymbol{\theta}_i^c$. This leads to scale-dependent random walk algorithm. In this case, α is of the form (4.3) with $J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i^c) = N(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c, \Sigma_{\boldsymbol{\theta}_i^c})$. Note that in this case the jumping distributions no longer symmetric in its arguments. Consequently the jumping distributions do not cancel as they do in Algorithm II. It is also the case that the discussion in Gelman et al. (1995b) applies.

Algorithm IV: Posterior approximation M-H algorithm

Finally, we can use a jumping distribution constructed to provide a good approxi-

mation to the target distribution. For example, $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c) = N(\boldsymbol{\theta}_i \mid \tilde{\boldsymbol{\theta}}_i, \Sigma_{\tilde{\boldsymbol{\theta}}_i})$, where $\tilde{\boldsymbol{\theta}}_i$ is a mode of the target distribution given $\boldsymbol{\theta}_i^c$. Alternatively the mean and variance of the jumping distribution can be derived based on a linear approximation to a nonlinear function. Since the jumping distribution is in this case an approximation to the target distribution, α for the traditional M-H algorithm is of the same form (4.3), with $J(\boldsymbol{\theta}_i^c \mid \boldsymbol{\theta}_i^*) = N(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\theta}}_i^*, \Sigma_{\boldsymbol{\theta}_i^*})$ and $J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i^c) = N(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\theta}}_i^c, \Sigma_{\boldsymbol{\theta}_i^c})$, where two different linearizations are carried out one at $\boldsymbol{\theta}_i^*$ and one at $\boldsymbol{\theta}_i^c$.

No matter what kind of jumping distribution is used for a M-H algorithm, there remains the question of how to choose the scale matrix $\boldsymbol{\Sigma}$. Gelman et al. (1995a, 1995b) suggest adjusting the scale of $\boldsymbol{\Sigma}$ in the random walk M-H to bring the acceptance rate between 0.23 to 0.45. They find that if $\boldsymbol{\Sigma}$ is the target posterior variance (usually unknown), then $2.4^2\boldsymbol{\Sigma}/d$ works well in the jumping distribution, where d is the dimension of $\boldsymbol{\Sigma}$.

When linearization of the nonlinear function is used to generate the jumping distribution, the variance matrix for the normal jumping distribution can be derived either from the Hessian matrix $\mathbf{H}$ or from the gradient vector $\mathbf{F}$. Vector $\mathbf{F}$ and matrix $\mathbf{H}$, respectively, are the first and the second derivatives of the relevant portion of the log-likelihood function with respect to the corresponding parameters. The variance can be taken as the inverse of the negative of the Hessian matrix or the inverse of the product $\mathbf{F}\mathbf{F}^T$. When the likelihood has a non-Gaussian shape at some points, the third derivative may be large at those points, which means the Hessian matrix may differ significantly from the Hessian matrix evaluated at nearby points (Thisted, 1989). Such instability can be a problem.

4.5 Batching and other issues

The way in which the model parameters $\boldsymbol{\theta}$ are partitioned in developing an MCMC algorithm may affect the convergence rate (Gilks et al., 1996). This is especially relevant

for the Gibbs sampler. For example, the fixed effects in the linear mixed model can be partitioned as $\mathbf{b} = (\mathbf{b_1} \ldots \mathbf{b_k} \ldots \mathbf{b_m})$ with each subvector corresponding to a subpopulation defined by the level of the fixed effects (e.g., gender), or they can be partitioned element by element as $\mathbf{b} = (b_{11}, b_{12}, \ldots, b_{1p}, b_{21}, \ldots, b_{mp})$. In the Markov chain simulation, the posterior sample is drawn in batches of size p for the former partition, while it is drawn element by element for the latter. If larger batches are used then one travels across the joint distribution with a multi-dimensional move to update all of the parameters in the group. For the single-element scheme, each move is one-dimensional. Since the elements within a parameter subvector may be correlated, the choice of partition influences the mixing of the Markov chain (Gilks et al., 1996; Gilks and Roberts, 1996). Hence, there is a need to study the impact of different batching algorithms in order to develop a good computational scheme.

Another issue that can affect the efficiency of the MCMC algorithm is the sensitivity to starting values. Starting values for MCMC algorithms do not of course affect the stationary distribution, assuming one exists. However, some algorithms appear to be more sensitive to the choice of starting values, working poorly (or not at all) for some choices and well for others. There are conflicting goals in selecting starting values. Starting values must be overdispersed with respect to the target distribution for the Gelman and Rubin convergence diagnosis (PSR) to be valid. However, starting values may need to be chosen extremely carefully for slow-mixing chains, so that an unusual starting value does not increase the time required for convergence (Gilks et al., 1996).

CHAPTER 5 LINEAR RANDOM REGRESSION MODELS

5.1 Introduction

Previous chapters introduced models that can be used to analyze phenotypic trait data measured longitudinally, reviewed likelihood and Bayesian inference, and discussed approaches that can improve the efficiency of MCMC algorithms for carrying out the Bayesian approach. The current chapter focuses on Bayesian inference for linear random regression models. Building on the discussion of Chapter 4, we provide empirical results concerning the efficiency of a number of MCMC algorithms with the goal of providing practical advice to users of those models.

To review, random regression (RR) models for longitudinal data incorporate population average response patterns over time and individual-specific effects. Individual response curves over time vary around the population average. Allowing individual variation enables one to accommodate animal genetic effects and permanent environmental effects, as well as individual-level covariates. In this chapter, we focus on the case when the population curve and individual effects are linear models (i.e., linear in the parameters). Chapter 6 addresses nonlinear random regression models.

We focus on the Bayesian approach to data analysis using RR models. The posterior distribution of model parameters is proportional to the product of the prior distribution and the sampling distribution, as described in Chapter 3. Bayesian methods rely on Markov Chain Monte Carlo (MCMC) simulation. It is obviously beneficial to develop and adopt efficient simulation algorithms to explore the posterior distributions. Factors

associated with efficiency of MCMC were described in Chapter 4. Here we study those factors in the context of RR models.

The random regression models used in this chapter are described in Section 5.2. Notation for the models is developed and complete Bayesian model specifications are given. The use of hierarchical centering and orthogonal polynomials for improving the efficiency of MCMC algorithms are discussed in Sections 5.3 and 5.4. An application of the polynomial random regression model to pig weight data is illustrated in Section 5.5. Results of the data analysis are provided but the focus is on the efficiency of the various MCMC algorithms. All models used in this chapter are listed in Table 5.1, and the notation for this chapter is listed in Table 5.2.

5.2 Random polynomial regression models

5.2.1 Model specification – independent animals

To begin we suppose repeated measurements of a growth trait are taken over time on n unrelated animals. Let r_i denote the number of records for animal i and let $\mathbf{y}_i$ be the corresponding $r_i \times 1$ vector of measurements. We often stack the individual vectors together with $\mathbf{y} = (\mathbf{y}_1^T\ \mathbf{y}_2^T\ \ldots\ \mathbf{y}_n^T)^T$, the $N \times 1$ vector of measurements, where $N = \sum_1^n r_i$. A number of RR models may be considered for a given data set. We let Mpq represent the random polynomial regression models with a p^{th} degree polynomial for the population (fixed effects) growth curve and a q^{th} ($q \leq p$) degree polynomial for individual-level effects (random effects) around the population average. This means that the first q polynomial coefficients vary from animal to animal due to individual effects.

In practice it may be desirable to allow separate population curves for each of m subpopulations (e.g., $m = 2$ with males and females serving as subpopulations). Let $\mathbf{b} = (\mathbf{b}_1\ \mathbf{b}_2 \ldots\ \mathbf{b}_m)$ denote the vector of subpopulation parameters, $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_n)$ denote the individual level parameter vector, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1\ \boldsymbol{\epsilon}_2\ \ldots\ \boldsymbol{\epsilon}_n)$ denote the ran-

Table 5.1 Abbreviations and brief description for polynomial random regression models used in Chapter 5

Abbreviation	Model Description
Mpq	$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}$ a random regression(RR) model with a p^{th} degree polynomial for fixed effects and a q^{th} degree polynomial for random effects, where $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G})$ and $\epsilon \sim N(0, \sigma_e^2)$
MpqL	model Mpq with Legendre polynomials
MpqR	$\mathbf{y} = \mathbf{X}^+\mathbf{b}^+ + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ a reparameterization of model Mpq based on hierarchical centering, with a q^{th} degree polynomial for random effects and a $(q+1)^{th}$ to p^{th} polynomial for fixed effects, where $\boldsymbol{\eta}_i \sim N(\mathbf{b}_1, \mathbf{G}))$ and $\epsilon \sim N(0, \sigma_e^2)$
MpqRL	model Mpq with Legendre polynomials
MpqA	$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zp} + \boldsymbol{\epsilon}$ an extension of model Mpq when genetic relationships among animals are incorporated, where $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A} \otimes \mathbf{G})$ and $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{E})$
MpqRA	a reparameterization of model MpqA based on hierarchical centering
MpqRAL	Model MpqRA with Legendre polynomials

Table 5.2 Notation used for linear random regression models

Abbreviation	Description
r_i	the number of repeated records on animal i
n_k	the number of animals in subpopulation k
n	the total number of animals in the data set. $n = \sum_{k=1}^{m} n_k$
N	the total number of records. $N=\sum_{i=1}^{n} r_i$
σ_e^2	the variance for random sampling residuals
$\mathbf{G}$	the variance matrix for the individual random effect parameters
$\mathbf{b}_k$	vector of the fixed effect parameters for the k^{th} subpopulation
$\mathbf{u}_i$	vector of the i^{th} animal's random effect parameters
$\mathbf{a}_i$	vector of the i^{th} animal genetic effect parameters
$\mathbf{p}_i$	vector of the i^{th} animal's permanent environmental effect parameters
$\boldsymbol{\eta}_i$	vector of the i^{th} animal's centered random effect parameters
$\boldsymbol{\eta}$	vector of centered random effect parameters
$\tilde{\mathbf{b}}_{k(i)}$	mean vector for $\boldsymbol{\eta}_i$
$\tilde{\mathbf{b}}$	mean vector for $\boldsymbol{\eta}$
$\mathbf{b}$	vector of fixed effect parameters
$\mathbf{u}$	vector of animal random effect parameters
$\mathbf{a}$	vector of animal genetic effect parameters
$\mathbf{p}$	vector of permanent environmental effect parameters
$\mathbf{X}_i$ or $\mathbf{Z}_i$	the incidence matrix corresponding to the i^{th} animal's fixed or random effect parameters
$\mathbf{A}$	the additive genetic relationship matrix between animals with entries denoted by A_{ij}
$\mathbf{A}^{-1}$	the inverse of matrix $\mathbf{A}$ with entries denoted by A^{ij}
MPSR	a scalar used for convergence diagnosis, summarizing the distance of between-sequence to within-sequence variance matrices of multiple parameters. MPSR approaches 1.0 at convergence
$\sqrt{\hat{R}^p}$	the estimate of MPSR
γ	convergence point: the number of iterations required to detect convergence of the Markov chain, including the burn-in

dom residual vector. The vectors $\mathbf{u}$ and ϵ contribute individual measurement variation above or below the fixed effects. Note that each vector $\mathbf{b}_k$ of subpopulation fixed effects is of length p: $\mathbf{b}_K = (b_{0K}\ b_{1K}\ \ldots b_{pK})^T$, and each vector $\mathbf{u}_i$ is of length q: $\mathbf{u}_i = (u_{0i}\ u_{1i}\ \ldots u_{qi})^T$. The model for measurements $\mathbf{y}$ can be written as

$$\begin{aligned}
\mathbf{y} &= \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_n^T \end{pmatrix}^T \\
&= \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \\
&= \mathbf{Xb} + \mathbf{Zu} + \epsilon, \qquad (5.1)
\end{aligned}$$

where $\epsilon \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$, $\mathbf{X}_i$ $(r_i \times m(p+1))$ and $\mathbf{Z}_i$ $(r_i \times (q+1))$ are incidence matrices associated with the polynomials in time for fixed effects $\mathbf{b}$ and random effects $\mathbf{u}_i$, respectively. Matrix $\mathbf{Z}_i$ includes $q+1$ columns corresponding to constant, linear, ... and q^{th} order polynomial terms in time. Matrix $\mathbf{X}_i$ is $\mathbf{0}$ in all columns except columns corresponding to the $p+1$ polynomial terms for the relevant subpopulation.

It is common to assume independent identical Gaussian distributions for the error terms in ϵ, so the likelihood function for the observations on n independent animals is

$$\begin{aligned}
L(\mathbf{b}, \mathbf{u}, \sigma_e^2 \mid \mathbf{y}) &= (2\pi\sigma_e^2)^{-0.5N} exp\left(-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})^T(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})\right) \qquad (5.2) \\
&= (2\pi\sigma_e^2)^{-0.5N} exp\left(-\frac{1}{2\sigma_e^2}\sum_i^n(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i)^T(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i)\right).
\end{aligned}$$

Assuming that random effects follow independent identical Gaussian distributions, taken as $\mathbf{u}_i \mid \mathbf{G} \sim N(\mathbf{0}, \mathbf{G})$, prior distributions for the remaining parameters for model Mpq are

- $\sigma_e^2 \sim I\chi^2(\nu_e, \sigma_0^2)$
- $\mathbf{G} \sim IW(\nu_g, \mathbf{G}_0^{-1})$

- $p(\mathbf{b}) \propto$ constant

where $I\chi^2$ denotes the inverse chi-square distribution with density

$$I\chi^2(\theta) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^{\nu} \theta^{-(\nu/2+1)} e^{-\nu s^2/(2\theta)}. \quad \theta > 0.$$

$\mathrm{E}(\theta) = \frac{\nu}{\nu-2} s^2$ and $var(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)} s^4$. IW denotes the inverse Wishart distribution with density

$$IW(\mathbf{W}) = \left(2^{\nu k/2} \pi^{k(k-1)/4} \Pi_{i=1}^{k} \Gamma(\frac{\nu+1-i}{2})\right)^{-1} |\mathbf{S}|^{\nu/2} |\mathbf{W}|^{-(\nu+k+1)/2} exp(\frac{-1}{2} tr(\mathbf{S}\mathbf{W}^{-1})).$$

where $\mathbf{W}$ is positive definite, $\mathrm{E}(\mathbf{W})=(\nu - k - 1)^{-1}\mathbf{S}$, and k is the dimension of $\mathbf{W}$. The degrees of freedom ν_e and ν_g are hyperparameters that control the contribution of the prior distribution to the posterior distribution. Small values for the degrees of freedom indicate large prior variance and greater weight on the data. The remaining hyperparameters σ_0^2 and $\mathbf{G}_0$ may be selected according to any prior information, e.g., estimates obtained from previous experiments on similar populations.

Given the likelihood function and the prior distributions, the joint posterior distribution of all parameters for model Mpq (5.1) is

$$\begin{aligned} p(\mathbf{b}, \mathbf{u}, \mathbf{G}, \sigma_e^2 \mid \mathbf{y}) &= p(\mathbf{y} \mid \mathbf{b}, \mathbf{u}, \sigma_e^2) p(\mathbf{b}) p(\mathbf{u} \mid \mathbf{G}) p(\mathbf{G}) p(\sigma_e^2) / p(\mathbf{y}) \\ &= p(\mathbf{b}) p(\mathbf{G}) p(\sigma_e^2) \prod_{i}^{n} p(\mathbf{y}_i \mid \mathbf{b}, \mathbf{u}_i, \sigma_e^2) p(\mathbf{u}_i \mid \mathbf{G}) / p(\mathbf{y}) \qquad (5.3) \end{aligned}$$

In the subsequent sections, we apply Gibbs sampling to generate values from the joint posterior distribution. The full conditional posterior distributions of the parameters that are required for Gibbs sampling are as follows.

- $\sigma_e^2 \sim I\chi^2 \left(\nu_e + N, \left(\nu_e \sigma_0^2 + \sum_i^n (\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i)^T (\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i)\right) / (\nu_e + N)\right)$.

- $\mathbf{G} \sim IW(\nu_g + n, (\mathbf{G_0} + \mathbf{S})^{-1})$, where $\mathbf{S} = \mathbf{U}^T\mathbf{U}$, and

$$\mathbf{U} = (\mathbf{U}_0, \mathbf{U}_1, \ldots, \mathbf{U}_q) = \begin{pmatrix} u_{01} & u_{11} & \cdots & u_{q1} \\ u_{02} & u_{12} & \cdots & u_{q2} \\ \vdots & \vdots & & \vdots \\ u_{0n} & u_{1n} & \cdots & u_{qn} \end{pmatrix}.$$

  The column vector $\mathbf{U}_j = (u_{j1}, u_{j2}, \ldots, u_{jn})^T$ is the vector of j^{th} degree coefficients for the polynomial of random effects for all animals.

- $\mathbf{b} \sim N(\hat{\mathbf{b}}, (\mathbf{XX})^{-1}\sigma_e^2)$. Note that by the definition of $\mathbf{X}_i$, the full conditional posterior distribution of $\mathbf{b}$ factors as the product of m independent Gaussian distributions with $\mathbf{b}_k = (b_{0k}\ b_{1k}\ \ldots b_{pk})^T \sim N(\hat{\mathbf{b}}_k, (\mathbf{X}^T_{k(i)=k}\mathbf{X}_{k(i)=k})^{-1}\sigma_e^2)$, where $k = 1, 2, \ldots, m$) indicates the subpopulation, $\hat{\mathbf{b}}_k = (\mathbf{X}^T_{k(i)=k}\mathbf{X}_{k(i)=k})^{-1}\mathbf{X}^T_{k(i)=k}(\mathbf{y}_{k(i)=k} - \mathbf{Z}_{k(i)=k}\mathbf{u}_{k(i)=k})$, and $k(i) = k$ indicates the rows and the columns corresponding to the k^{th} subpopulation. For example, $\mathbf{X}_{k(i)=k}$ is the collection of incidence matrices $\mathbf{X}_i$ in $\mathbf{X}$ in (5.1) for those animals associated with the k^{th} subpopulation.

- $\mathbf{u_i} = (u_{0i}, u_{1i}, \ldots, u_{qi})^T \sim N(\hat{\mathbf{u}}_i, \hat{\mathbf{\Sigma}}_{u_i})$, where $\hat{\mathbf{\Sigma}}_{u_i} = (\mathbf{Z}_i^T\mathbf{Z}_i/\sigma_e^2 + \mathbf{G}^{-1})^{-1}$, and $\hat{\mathbf{u}}_i = \hat{\mathbf{\Sigma}}_{u_i}\mathbf{Z}_i^T(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{k(i)})/\sigma_e^2$.

Since all conditional posterior distribution are standard distributions, it is straightforward to implement the Gibbs sampler to draw samples from the posterior distribution.

5.2.2 Incorporating the relationship between animals

The model in the previous section assumed unrelated animals. In practice the animals available for study are typically related. In that case, it is natural to divide the individual random effects into genetic and permanent environmental effects. The model with a p^{th} degree polynomial for fixed effects and a q^{th} degree polynomial for random effects that incorporates animal genetic relationship is named model MpqA. The key idea is

to write $\mathbf{u}_i$ as $\mathbf{a}_i + \mathbf{p}_i$ where $\mathbf{a}_i = (a_{0i}\ a_{1i}\ \ldots a_{qi})^T$ are additive genetic effects, and $\mathbf{p}_i = (p_{0i}\ p_{1i}\ \ldots p_{qi})^T$ are permanent environmental effects. If we take $\mathbf{a} = (\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_n)$ and $\mathbf{p} = (\mathbf{p}_1\ \mathbf{p}_2\ \ldots\ \mathbf{p}_n)$, then the model for the response $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zp} + \boldsymbol{\epsilon}. \tag{5.4}$$

It is common to assume that $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}_a)$ and $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{E})$, where $\mathbf{A}$ is the additive genetic relationship matrix among animals with element $\mathbf{A}_{ij}$, and $\otimes$ denotes the Kronecker product. We illustrate the use of the Kronecker product by explaining $var(\mathbf{a}) = \mathbf{A} \otimes \mathbf{G}_a$. Matrix $\mathbf{A}$ is of dimension $n \times n$ and $\mathbf{G}_a$ is $(q+1) \times (q+1)$, then $\mathbf{A} \otimes \mathbf{G}_a$ is $n(q+1) \times n(q+1)$, and can be written in block form as

$$\begin{pmatrix} A_{11}\mathbf{G}_a & A_{12}\mathbf{G}_a & \ldots & A_{1n}\mathbf{G}_a \\ A_{12}\mathbf{G}_a & A_{22}\mathbf{G}_a & \ldots & A_{2n}\mathbf{G}_a \\ \vdots & \vdots & \vdots & \vdots \\ A_{1n}\mathbf{G}_a & A_{2n}\mathbf{G}_a & \ldots & A_{nn}\mathbf{G}_a \end{pmatrix}$$

With this extended model, the remaining development of Section 5.2.1 remains virtually changed. We do require an additional prior distribution for $\mathbf{E}$ and assume that $\mathbf{E} \sim IW(\nu_p, \mathbf{E}_0^{-1})$. Then the full conditional posterior distribution for $\mathbf{E}$ is $\mathbf{E} \sim IW(\nu_p + n, (\mathbf{E}_0 + \mathbf{S})^{-1})$, where $\mathbf{S} = \mathbf{P}^T\mathbf{P}$. Matrix $\mathbf{P}$ is of the same form as $\mathbf{U}$ but substituting u_{ji} with p_{ji}. As the distribution for $\mathbf{a}$ is different than that for $\mathbf{u}$, the full conditional posterior distributions for $\mathbf{a}$ and $\mathbf{p}$ are modified as follows.

- $\mathbf{a}_i \sim N(\hat{\mathbf{a}}_i, \hat{\boldsymbol{\Sigma}}_{\mathbf{a}_i})$, where $\hat{\boldsymbol{\Sigma}}_{\mathbf{a}_i} = (A^{ii}\mathbf{G}^{-1} + \mathbf{Z}_i^T\mathbf{Z}_i/\sigma_e^2)^{-1}$,
  $\hat{\mathbf{a}}_i = \hat{\boldsymbol{\Sigma}}_{\mathbf{a}_i}\left(\mathbf{Z}_i^T(\mathbf{y} - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{p}_i)/\sigma_e^2 - \sum_{j \neq i}^{n} A^{ij}\mathbf{G}^{-1}\mathbf{a}_j\right)$, where A^{ij} is the ij^{th} entry of $\mathbf{A}^{-1}$.
- $\mathbf{p}_i \sim N(\hat{\mathbf{p}}_i, \hat{\boldsymbol{\Sigma}}_{\mathbf{p}_i})$, where $\hat{\boldsymbol{\Sigma}}_{p_i} = (\mathbf{Z}_i^T\mathbf{Z}_i/\sigma_e^2 + \mathbf{E}^{-1})^{-1}$, and
  $\hat{\mathbf{p}}_i = \hat{\boldsymbol{\Sigma}}_{p_i}\mathbf{Z}_i^T(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{a}_i)/\sigma_e^2$.

5.2.3 Convergence rate

An iterative simulation is said to have converged when the parallel Markov chains are indistinguishable for any quantity of interest. Figure 5.1 shows a case where lack of convergence (in the first 200 iterations) is evident from comparing parallel sequences. Notice that convergence can sometimes not be detected from one sequence alone (e.g., (b) in Figure 5.1)(Gelman, 1996). To diminish the effect of the starting points, convergence diagnosis and inference focus on the second halves of the Markov chains. The first half is used as burn-in.

The parameter space is multi-dimensional for the models under study. The multivariate potential scale reduction (MPSR) criterion introduced by Brooks and Gelman (1998) is a quantitative approach for convergence diagnosis (see Section 3.3.4). MPSR is based on analysis of variance methods. Approximate convergence is diagnosed when the between-sequence variance is no larger than the within-sequence variance. Hence, when MPSR is near 1, we conclude that the simulated observations are from the target distribution. An estimate of MPSR is denoted by $\sqrt{\hat{R}^p}$ (see Section 3.3.4.3), and an estimated value of MPSR below 1.2 is generally considered as indicating convergence.

In practice $\sqrt{\hat{R}^p}$ values are calculated every s iterations during the Markov chain simulation. In our study, a simulation process is diagnosed as being converged at iteration γ, when $\sqrt{\hat{R}^p}$ remains below 1.2 for at least 20,000 iterations afterward. Then we use the length γ (including burn-in) to indicate the convergence point. The faster the simulation converges, the smaller γ will be.

5.3 Hierarchical centering

Hierarchical centering (see Section 4.2) is a strategy used to yield lower correlations between blocks of parameters and better mixing of MCMC simulations (Gelfand et al., 1995a; Gilks and Roberts, 1996). In the traditional parameterization, random effects

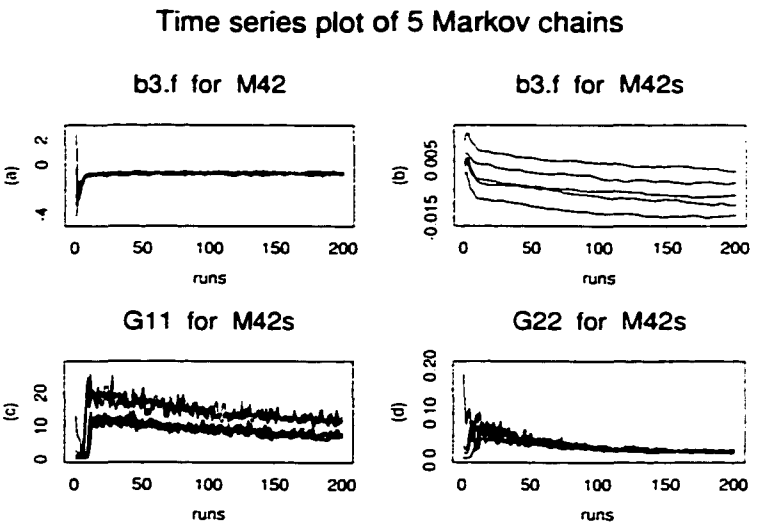

Figure 5.1 Time series plot of first 200 iterations to show the mixing speed of Markov chains

represent the deviation of an individual's model from the population average model, and are often assumed to be normally distributed with mean zero. Hierarchical centering provides an equivalent model specification in which the random effects enter as part of a hierarchy, data depend on random effect parameters, and random effect parameters are modeled as random draws from a model centered at the population-level parameters (rather than zeros).

5.3.1 Hierarchical centering for random regression models

For the polynomial random regression models that we consider, polynomials of different degree are fitted to the fixed and random effects. Model Mpq has a p^{th}-degree polynomial of time for fixed effects (a separate polynomial for each of m subpopulations is permitted) and a q^{th}-degree polynomial of time for random effects.

When the degree of the polynomial for random effects is less than that for fixed effects ($p > q$), the centered model MpqR is derived by centering all random effects at the means of the corresponding fixed effects instead of at zeros. Model MpqR then takes the form

$$\mathbf{y} = \mathbf{X}^{+}\mathbf{b}^{+} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}. \qquad \text{(model } \mathbf{MpqR}\text{)} \tag{5.5}$$

where $\mathbf{X}^{+}$ is the incidence matrix containing the higher order polynomial terms (degree $q+1$ to p) that are included as fixed effects but not as random effects. $\mathbf{b}^{+} = (\mathbf{b}_1^{+} \ldots \mathbf{b}_m^{+})$ are the fixed effect parameters corresponding to higher degree polynomial terms. $\mathbf{b}_k^{+} = (b_{(q+1)k} \ \ldots \ b_{pk})^T$, and $\mathbf{Z}$ is the incidence matrix for the random effects (same as in (5.1)). The vector of the centered random effects for individual i, $\boldsymbol{\eta}_i = (\eta_{0i} \ \eta_{1i} \ \ldots \ \eta_{qi})^T$, is assumed to follow a Gaussian distribution $N(\tilde{\mathbf{b}}_k, \mathbf{G})$, where k is the subpopulation to which animal i belongs, and $\tilde{\mathbf{b}}_k = (b_{0k} \ b_{1k} \ldots b_{qk})$ are the corresponding fixed effect parameters for the q degree polynomial. Note that the centered random effects can be expressed in terms of the parameters of the uncentered model as $\boldsymbol{\eta}_i = \tilde{\mathbf{b}}_k + \mathbf{u}_i$, where $\mathbf{u}_i$ are the random effects with mean zero.

We next consider how this parameterization affects the likelihood function and the full conditional posterior distributions that are used in MCMC algorithms. Assuming n animals are unrelated, the likelihood function is

$$\begin{aligned} L\left(\mathbf{b}, \boldsymbol{\eta}, \sigma_e^2 \mid \mathbf{y}\right) &= L\left(\mathbf{b} = (\tilde{\mathbf{b}}, \mathbf{b}^{+}), \boldsymbol{\eta}, \sigma_e^2 \mid \mathbf{y}\right) \\ &= (2\pi\sigma_e^2)^{-0.5N} exp\left(-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{X}^{+}\mathbf{b}^{+} - \mathbf{Z}\boldsymbol{\eta})^T(\mathbf{y} - \mathbf{X}^{+}\mathbf{b}^{+} - \mathbf{Z}\boldsymbol{\eta})\right) \\ &= (2\pi\sigma_e^2)^{-0.5N} exp\left(-\frac{1}{2\sigma_e^2}\sum_i^n(\mathbf{y}_i - \mathbf{X}_i^{+}\mathbf{b}^{+} - \mathbf{Z}_i\boldsymbol{\eta}_i)^T(\mathbf{y}_i - \mathbf{X}_i^{+}\mathbf{b}^{+} - \mathbf{Z}_i\boldsymbol{\eta}_i)\right). \end{aligned}$$

The remaining prior distributions are set up as follows.

- $\sigma_e^2 \sim I\chi^2(\nu_e, \sigma_0^2)$
- $\mathbf{G} \sim IW(\nu_g, G_0^{-1})$

- $p(\mathbf{b}^+) \propto$ constant, $p(\tilde{\mathbf{b}}) \propto$ constant

Given the likelihood functions and the prior distributions , the joint posterior distribution of all parameters for model MpqR (5.5) is

$$p(\tilde{\mathbf{b}}, \mathbf{b}^+, \boldsymbol{\eta}, \mathbf{G}, \sigma_e^2 \mid \mathbf{y}) = p(\mathbf{y} \mid \mathbf{b}^+, \boldsymbol{\eta}, \sigma_e^2) p(\mathbf{b}^+) p(\boldsymbol{\eta} \mid \tilde{\mathbf{b}}, \mathbf{G}) p(\tilde{\mathbf{b}}) p(\mathbf{G}) p(\sigma_e^2) / p(\mathbf{y})$$

The full conditional posterior distributions are as follows:

- $\sigma_e^2 \sim I\chi^2\left(\nu_e + N, \left(\nu_e \sigma_0^2 + (\mathbf{y} - \mathbf{X}^+\mathbf{b}^+ - \mathbf{Z}\boldsymbol{\eta})^T(\mathbf{y} - \mathbf{X}^+\mathbf{b}^+ - \mathbf{Z}\boldsymbol{\eta})\right)/(\nu_e + N)\right)$.
- $\mathbf{G} \sim IW(\nu_g + n, (\mathbf{G_0} + \mathbf{S})^{-1})$, where $\mathbf{S} = \mathbf{T}^T\mathbf{T}$, $\mathbf{T} = (\mathbf{T}_0, \mathbf{T}_1, \ldots, \mathbf{T}_q)$ with $\mathbf{T}_j = (\eta_{j1}^* \ \eta_{j2}^* \ \ldots \ \eta_{jn}^*)^T$; $j = 0, 1, \ldots, q$, and $\eta_{ji}^* = \eta_{ji} - \tilde{b}_{jk}$, if animal i belongs to subpopulation k.
- $\mathbf{b}_k^+ \sim N(\hat{\mathbf{b}}_k^+, (\mathbf{X}_{k(i)=k}^{+T}\mathbf{X}_{k(i)=k}^+)^{-1}\sigma_e^2)$, where $\hat{\mathbf{b}}_k^+ = (\mathbf{X}_{k(i)=k}^{+T}\mathbf{X}_{k(i)=k}^+)^{-1}\mathbf{X}_{k(i)=k}^{+T}(\mathbf{y}_{k(i)=k} - \mathbf{Z}_{k(i)=k}\boldsymbol{\eta}_{k(i)=k})$ and $k(i) = k$ indicates the rows and columns corresponding to subpopulation k.
- $\tilde{\mathbf{b}}_k \sim N(\bar{\boldsymbol{\eta}}_k, \mathbf{G}/n_k)$, where n_k is the total number of animals in the k^{th} subpopulation, and $\bar{\boldsymbol{\eta}}_k = \sum_{k(i)=k} \boldsymbol{\eta}_i / n_k$ is the average of the random effects for level k.
- $\boldsymbol{\eta}_i \mid \mathbf{b}_1, \mathbf{G} \sim N(\hat{\boldsymbol{\eta}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_i}\sigma_e^2)$, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_i} = (\mathbf{Z}_i^T\mathbf{Z}_i/\sigma_e^2 + \mathbf{G}^{-1})^{-1}$ and

  $$\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}_i}\left(\mathbf{Z}_i^T(\mathbf{y}_i - \mathbf{X}_i^+\mathbf{b}^+)/\sigma_e^2 + \mathbf{G}^{-1}\tilde{\mathbf{b}}_k\right);$$

  if animal i belongs to the k^{th} subpopulation.

One case that deserves special attention is the case when $p = q$. In that case, the model (MqqR) can be written as $\mathbf{y} = \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\eta}_i \sim N(\tilde{\mathbf{b}}_k, \mathbf{G})$, where k indicates the subpopulation that animal i belongs to. This avoids some of the awkward notation that is required to address the split of $\mathbf{b}_k$ into two subvectors. The full conditional posterior distribution for $\tilde{\mathbf{b}}_k$ is of the same form as $\tilde{\mathbf{b}}_k$ for model MpqR.

The version of the model that incorporates genetic and permanent environmental effects in place of the single vector of random effects can also be modified to include hierarchical centering. Recall that the form of the model is $\mathbf{y}=\mathbf{Xb}+\mathbf{Za}+\mathbf{Zp}+\epsilon$. There are two options for hierarchical centering: either centering genetic effects (thus using parameters ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{p}$) and $\boldsymbol{\eta} = \tilde{\mathbf{b}} + \mathbf{a}$) or centering permanent environmental effects (using parameters ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{a}$) and $\boldsymbol{\eta} = \tilde{\mathbf{b}} + \mathbf{p}$), where $\tilde{\mathbf{b}}$ represents fixed effects with the same degree of polynomial as those used for the random effects $\mathbf{a}$ or $\mathbf{p}$. The decision as to whether to center or not and how to center are, therefore, more complex for model MpqA. We will come back to this issue in the next section.

5.3.2 When is hierarchical centering preferred ?

Gelfand et al. (1995) provide results to guide the choice between models Mpq and MpqR. Their result (see Section 4.2.2) shows that the more efficient MCMC algorithm depends on the relative magnitude of the variance of the conditional posterior distribution of the centered parameters $\boldsymbol{\eta}_i$ (given above as $(\mathbf{Z}_i^T\mathbf{Z}_i/\sigma_\epsilon^2 + \mathbf{G}^{-1})^{-1}$) to the variance of uncentered parameters $\mathbf{u}_i$ (which is $\mathbf{G}$). The hierarchically centered model will improve the convergence of MCMC algorithms, if the ratio of these two variances is close to zero. This criterion indicates that centering will be effective when $\left|\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{G}/\sigma_\epsilon^2 + \mathbf{I}\right|^{-1}$ is close to zero. Essentially this occurs when the variance of the random effects, $\mathbf{G}$, is large compared to σ_ϵ^2.

For model MpqA, the optimal parameterization depends on the relative magnitude of variance matrices $\mathbf{G}$, $\mathbf{E}$ and σ_ϵ^2. In practice, it is not obvious which parameterization to select, since $\mathbf{G}$ and $\mathbf{E}$ are unknown in advance. Even when the relative magnitude of the two variance matrices is known, the choice of parameterization can be affected by the sample size or by the degree of relationship between animals (given by $\mathbf{A}$). In this context, a cycling algorithm is recommended by Gelfand and Carlin (1995). The cycling algorithm combines two or more algorithms into a single MCMC algorithm. One full

cycle consists of a single iteration for each of the component MCMC algorithms. Gelfand and Carlin find that the efficiency of the cycling algorithm is approximately equivalent to that of the best component. Therefore, in the context of our model, there are three possible MCMC implementations for model MpqA: (1) centering genetic effects (**b** $\boldsymbol{\eta}$ **p**), (2) centering permanent environmental effects (**b** $\boldsymbol{\eta}$ **a**), and (3) cycling the two centered parameterizations (**b** $\boldsymbol{\eta}$ **p**) and (**b** $\boldsymbol{\eta}$ **a**) in sequence. It is also possible to include the original uncentered parameters in the cycling algorithm but we have not usually found that useful in the models we consider.

5.4 Orthogonal polynomials

Correlations among parameters tend to slow the mixing of Gibbs sampling chains (Gilks and Roberts, 1996). Orthogonal transformation of the incidence matrices in polynomial regression models is an approach for reducing the correlation between the regression coefficients, which, therefore, could improve the convergence rate for the MCMC algorithm.

5.4.1 Legendre polynomials

There are a number of families of orthogonal polynomials (see Section 4.3) that can be used. We focus on the Legendre polynomials (Thisted, 1988). Legendre polynomials are continuous, normalized and orthogonal on (-1, 1), and defined as the following,

$$\begin{aligned} o_0(x) &= \sqrt{\frac{1}{2}} \\ o_1(x) &= \sqrt{\frac{3}{2}}x \\ o_k(x) &= \sqrt{\frac{2k+1}{2}}\frac{1}{2^k}\sum_{j=0}^{[k/2]}(-1)^j \begin{pmatrix} k \\ j \end{pmatrix} \begin{pmatrix} 2k-2j \\ k \end{pmatrix} x^{k-2j}, \end{aligned} \tag{5.6}$$

where k is the degree of the polynomial. They can be developed via a recurrence relationship (as described in Section 4.3) with $a_j = 0$, $b_j = (2j-1)/j$, and $c_j = (j-1)/j$ in (4.1). The recurrence relationship yields the unnormalized Legendre polynomials as

$$\begin{aligned} o_0^{un}(x) &= 1 \\ o_1^{un}(x) &= x \\ o_k^{un}(x) &= \frac{2k-1}{k}\, x\, o_{k-1}(x) - \frac{k-1}{k} o_{k-2}(x). \end{aligned} \tag{5.7}$$

These can then be normalized to have constant variance by multiplying the k^{th} polynomial by $\sqrt{(2k+1)/2}$. The Legendre polynomials are an even function of x when k is even, and an odd function when k is odd.

Let $\mathbf{\Lambda}_p$ be the matrix required to transform the first p monomials $(1\ x\ x^2\ \ldots\ x^p)$ into a p^{th} degree Legendre polynomial. For example, the first five Legendre polynomials are

$$\mathbf{\Lambda}_4 = \sqrt{0.5},\ \sqrt{1.5}x,\ \sqrt{2.5}(1.5x^2-0.5),\ \sqrt{3.5}(2.5x^3-1.5x),\ \text{and}\ \sqrt{4.5}(\frac{35}{8}x^4-\frac{15}{4}x^2+\frac{3}{8})$$

where $x \in (-1, 1)$. Then $\mathbf{\Lambda}_4$ is defined as

$$\mathbf{\Lambda}_4 = \begin{pmatrix} \sqrt{0.5} & 0 & -0.5\sqrt{2.5} & 0 & \frac{3}{8}\sqrt{4.5} \\ 0 & \sqrt{1.5} & 0 & -1.5\sqrt{3.5} & 0 \\ 0 & 0 & 1.5\sqrt{2.5} & 0 & -\frac{15}{4}\sqrt{4.5} \\ 0 & 0 & 0 & 2.5\sqrt{3.5} & 0 \\ 0 & 0 & 0 & 0 & \frac{35}{8}\sqrt{4.5} \end{pmatrix}.$$

so that row vector $o = (1\ t\ t^2\ t^3\ t^4)\mathbf{\Lambda}_4$ yields the appropriate quartic Legendre polynomial at time t.

Since the support for the Legendre Polynomials is on (-1,1), the time variable for the longitudinal data needs to be adjusted to match the support of the Legendre polynomial function. The adjusted time variable is calculated by

$$t^* = -1 + \frac{t - min(t)}{max(t) - min(t)} * 2.$$

where t is on the original time scale, t^* is the adjusted time on the (-1, 1) scale, and $min(t)$ and $max(t)$ are the minimum and maximum of time for the data at hand over all individuals. For example, if the measuring time varies from a minimum of 10 to a maximum of 90, then the adjusted time for t=30 is t^*=-1+(30-10)/(90-10)× 2=-0.5. For the day with adjusted time t^*=-.5, we can compute the first three orthogonal polynomial values as (0.7071, -0.6124, -0.1976). These values would then be used in the corresponding columns of $\mathbf{X}_i$ or $\mathbf{Z}_i$ in the random regression models.

5.4.2 Relationship between models Mpq and MpqL

Let MpqL and MpqRL denote the models we obtain by transforming to orthogonal polynomials in models Mpq and MpqR, respectively. Let $\mathbf{L}_p$ denote the transformation matrix between time on the regular scale and adjusted time on the (-1,1) scale for the p^{th} degree polynomial of time, so that $\tilde{\mathbf{X}}_i = \mathbf{X}_i\mathbf{L}_p^{-1}$ is on (-1,1) scale. $\mathbf{L}$ is data-dependent and model-dependent. Its structure will be illustrated in the subsequent section. Note that in this section $\mathbf{X}_i$ denotes the matrix with p columns and r_i rows for animal i. Suppose the n animals are independent (no relationship matrix is used) and animal i belongs to the k^{th} subpopulation (with fixed effects, vector $\mathbf{b}_k$), then the relationship between model Mpq (5.1) on the regular scale and model MpqL on the Legendre scale can be developed as follows.

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{X}_i\mathbf{b}_k + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i \\
&= \mathbf{X}_i\mathbf{L}_p^{-1}\boldsymbol{\Lambda}_p\boldsymbol{\Lambda}_p^{-1}\mathbf{L}_p\mathbf{b}_k + \mathbf{Z}_i\mathbf{L}_q^{-1}\boldsymbol{\Lambda}_q\boldsymbol{\Lambda}_q^{-1}\mathbf{L}_q\mathbf{u}_i + \boldsymbol{\epsilon}_i \\
&= \mathbf{X}_i^*\mathbf{b}_k^* + \mathbf{Z}_i^*\mathbf{u}_i^* + \boldsymbol{\epsilon}_i.
\end{aligned}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices on the original time scale (i.e., the entry for the element in the i^{th} row and k^{th} column is t_{ij}^{k-1}, see (2.2)), $\mathbf{X}_i^* = \mathbf{X}_i\mathbf{L}_p^{-1}\boldsymbol{\Lambda}_p$ and $\mathbf{Z}_i^* = \mathbf{Z}_i\mathbf{L}_q^{-1}\boldsymbol{\Lambda}_q$ are design matrices on the Legendre scale, and $\mathbf{b}_K^* = \boldsymbol{\Lambda}_p^{-1}\mathbf{L}_p\mathbf{b}_k$ and $\mathbf{u}_i^* = \boldsymbol{\Lambda}_q^{-1}\mathbf{L}_q\mathbf{u}_i$ are parameters on the Legendre scale. Note that the random effects variance

matrix $\mathbf{G}$ on the Legendre scale is

$$\mathbf{G}^* = var(\mathbf{u}_i^*) = \mathbf{\Lambda}_q^{-1}\mathbf{L}_q \, var(\mathbf{u}_i) \, (\mathbf{\Lambda}_q^{-1}\mathbf{L}_q)^T = \mathbf{\Lambda}_q^{-1}\mathbf{L}_q \, \mathbf{G} \, (\mathbf{\Lambda}_q^{-1}\mathbf{L}_q)^T$$

After the posterior distribution is obtained on the Legendre scale for model MpqL, parameters on the regular scale can be obtained by the reverse transformations $\mathbf{b}_k = \mathbf{L}_p^{-1}\mathbf{\Lambda}_p\mathbf{b}_k^*$, $\mathbf{u}_i = \mathbf{L}_q^{-1}\mathbf{\Lambda}_q\mathbf{u}_i^*$, and $\mathbf{G} = \mathbf{L}_q^{-1}\mathbf{\Lambda}_q\mathbf{G}^*\mathbf{\Lambda}_q^T\mathbf{L}_q^{-T}$.

5.4.3 Benefit of orthogonal polynomials

To understand why the use of orthogonal polynomials can be expected to produce more efficient MCMC algorithms, we consider the posterior distribution of the random effect parameters and fixed effects, given the variance components. Consider the structure of Henderson's mixed model equations (MME), given the variance parameters, which are expressed as follows for model Mpq.

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + (\mathbf{A}^{-1} \otimes \mathbf{G}^{-1}) \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{Y} \\ \mathbf{Z}^T\mathbf{Y} \end{pmatrix}. \tag{5.8}$$

where $\mathbf{A}$ is the additive genetic relationship matrix between animals and $\mathbf{I}$ will replace $\mathbf{A}$ when animals are independent. The variance for the BLUP estimates $(\hat{\mathbf{b}}\ \hat{\mathbf{u}})$, treating the variance parameters as fixed, is the inverse of the coefficient matrix in Henderson's MME. When $\mathbf{X}^T\mathbf{Z} \neq \mathbf{0}$, the fixed and random effect parameter estimates are correlated. The above is given in terms of the REML-BLUP approach. The same logic is relevant in the Gibbs sampling conditional distribution. Dependence between subsets of model parameters may slow the travel along the surface of the joint posterior distribution when using the Gibbs sampler (Gilks and Roberts, 1996).

Now we consider models MpqL and MpqRL (similar results are obtained for models MpqAL and MpqRAL). For model MpqL, the columns of the design matrix $\mathbf{Z}_i^*$ are a subset of $\mathbf{X}_i^*$ (the two matrices are equivalent if $p = q$). Assuming that animals 1 and n

belong to subpopulation 2 and animal 2 belongs to subpopulation 1, then the coefficient matrix for Henderson's MME for model MpqL is

$$\begin{pmatrix} \begin{pmatrix} 0 & I_p & \dots & 0 \\ I_p & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & I_p & \dots & 0 \end{pmatrix} & \begin{pmatrix} 0 & \begin{pmatrix} I_q \\ 0 \end{pmatrix} & \dots & 0 \\ \begin{pmatrix} I_q \\ 0 \end{pmatrix} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \begin{pmatrix} I_q \\ 0 \end{pmatrix} & \dots & 0 \end{pmatrix} \\ \text{symmetry} & I_{nq} + (A^{-1} \otimes G^{-1}) \end{pmatrix}$$

where $\mathbf{A} = \mathbf{I}$ when animal genetic relationships are ignored. The inverse of the above matrix is proportional to the variance of the conditional posterior distribution of model coefficients ($\mathbf{b}^*$ and $\mathbf{u}^*$, given the variance parameters). In this coefficient matrix, a significant number of off-diagonal entries are zeros, suggesting much lower posterior correlations between model coefficients.

For model MpqRL, $\mathbf{X}_i^{+*T}\mathbf{Z}_i^* = \mathbf{0}$ (with $\mathbf{X}_i^* = (\mathbf{Z}_i^* \ \mathbf{X}_i^{*+})$), so the coefficient matrix in Henderson's MME is

$$\begin{pmatrix} \begin{pmatrix} 0 & I_{p-q} & \dots & 0 \\ I_{p-q} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & I_{p-q} & \dots & 0 \end{pmatrix} & 0 \\ 0 & I_{nq} + (A^{-1} \otimes G^{-1}) \end{pmatrix}.$$

This matrix shows the posterior conditional independence between the fixed effects and random effects given the variance parameters . In addition, when fitting a model with the same degree of polynomial for fixed and random effects (say model Mqq), the model obtained by transforming to orthogonal polynomials (model MqqRL, $\mathbf{y} = \mathbf{Z}^*\boldsymbol{\eta}^* + \boldsymbol{\epsilon}$) provides independence between model coefficients, since $\mathbf{Z}^{*T}\mathbf{Z}^* = \mathbf{I}$.

Similar results as discussed above are obtained for models MpqAL and MpqRAL when relationships among animals are incorporated. However, close genetic relationships increase the dependence of random coefficients.

5.5 An application to pig weight gains

5.5.1 Data and model

The data set of pig weight gains used as an example in this study was obtained through the generosity of S. Andersen and B. Petersen of The National Committee for Pig Breeding, Health and Production in Denmark. A detailed description of the data set can be found in Andersen and Petersen (1996). Live weights of 190 slaughter pigs, 95 pigs of each gender, were measured from 4 weeks of age (a weight of about 25 kg) to approximately 20 weeks of age. Half of the pigs were slaughtered when their live weight reached 95 kg and the rest were slaughtered at 115 kg. All pigs were raised on the same farm. The number of times that weights were taken, r_i, varied from 19 to 32 across the 190 animals. All weights were slightly different at the start of test. Live weight gains (y) were recorded as 0 on the first day of test. The maximum number of days on test was 114. There were 37 families with 4 to 6 offspring per family, evenly divided among 2 to 3 sons and daughters. One family had only 1 son. The total number of records, $N = \sum_1^n r_i$, was 4294. Males and females were considered as separate subpopulations in the random regression models so that a separate population average growth curve was assumed for each gender. Figure 5.2 shows the weight gains of several animals over the test period. A common growth pattern is observed but presence of individual variation is also clear.

Treating the animals as unrelated, random regression model M42 was used in Andersen and Petersen's paper (1996). Under this model, the weight gain of the j^{th} measure

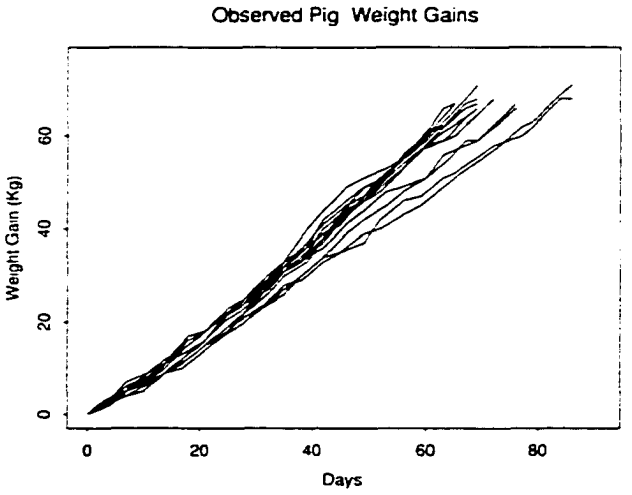

Figure 5.2 Observed weight gains from the start of test (day 0) of several pigs over time

for animal i, taken at time t_{ij} can be expressed as

$$\begin{aligned} y_{ij} &= b_{0k(i)} + b_{1k(i)}t_{ij} + b_{2k(i)}t_{ij}^2 + b_{3k(i)}t_{ij}^3 + b_{4k(i)}t_{ij}^4 + u_{0i} + u_{1i}t_{ij} + u_{2i}t_{ij}^2 + \epsilon_{ij} \\ &= \mathbf{t}_{4_{ij}}^T \mathbf{b}_{k(i)} + \mathbf{t}_{2_{ij}}^T \mathbf{u}_i + \epsilon_{ij}. \end{aligned} \tag{5.9}$$

where $\mathbf{b}_{k(i)} = (b_{0k(i)}\ b_{1k(i)}\ b_{2k(i)}\ b_{3k(i)}\ b_{4k(i)})^T$ are the fixed effect coefficients (population average for subpopulation k), $\mathbf{u}_i = (u_{0i}\ u_{1i}\ u_{2i})^T \sim N(\mathbf{0}, \mathbf{G})$ are the random effect coefficients for animal i, $\mathbf{t}_{p_{ij}} = (1\ t\ t^2 \ldots\ t^p)$ denotes a vector containing the 0^{th} through p^{th} order monomial in time, and $\epsilon_{ij} \sim N(0, \sigma_e^2)$. The model includes two vectors of fixed effects coefficients, one for males and the other for females (denoted as $k = M$ or F).

5.5.2 REML-BLUP results

Andersen and Petersen (AP, 1996) fit random regression model M42 to the pig weight data by a REML-BLUP approach. The REML-BLUP estimates for model M42 listed in Table 1 under AP are as follows, using subscript $REML$ to denote the estimates (for fixed and random effect parameters).

Fixed effects were:

$$\text{for boars:} \quad \hat{\mathbf{b}}_{M.REML} = \begin{pmatrix} 0.224 & 0.651 & 0.0108 & -0.00014 & 5.53\text{E-}7 \end{pmatrix}$$

$$\text{for gilts:} \quad \hat{\mathbf{b}}_{F.REML} = \begin{pmatrix} 0.302 & 0.628 & 0.0142 & -0.00018 & 7.29\text{E-}7 \end{pmatrix}.$$

and the REML estimates for residuals and var-covariances of random effects were

$$\hat{\sigma}^2_{\epsilon.REML} = 1.05, \qquad \hat{\mathbf{G}}_{REML} = var(\mathbf{u}_i) = \begin{pmatrix} 0.977 & -0.0632 & 4.32\text{E-}4 \\ -0.0632 & 0.0149 & -1.17\text{E-}4 \\ 4.32\text{E-}4 & -1.17\text{E-}4 & 1.68\text{E-}6 \end{pmatrix}.$$

Note that there are minor difference between males and females.

5.5.3 Bayesian analysis with independent animal model

5.5.3.1 Model M42 results

In order to compare with the results obtained by AP, the same M42 random polynomial regression model is adopted as a starting point in this study. We briefly review the essential elements of the Bayesian analysis here. More details are provided in Section 5.2.1.

Following the AP model, we ignore relationships between animals for the present time. The entire observation vector $\mathbf{y}$ can be written in the form (2.3),

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}$$

$$= \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \begin{pmatrix} \mathbf{b}_M \\ \mathbf{b}_F \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} + \boldsymbol{\epsilon} \qquad (5.10)$$

where the rows in $\mathbf{Z}_i$ are $\mathbf{t}_{2,j}^T$ and the rows in $\mathbf{X}_i$ are $\mathbf{t}_{4,j}^T$ in the column for the corresponding gender and zero elsewhere. The between-individual variation is introduced through $\mathbf{u}_i$ for this linear hierarchical model and $\mathbf{u}$ is assumed to follow $N(\mathbf{0}, \mathbf{I} \otimes \mathbf{G})$.

Prior distributions are selected as in Section 5.2.1 and are listed in Table 5.3 for model M42. For convenience, hyperparameters σ_0^2 or $\mathbf{G}_0$ are chosen so that the mean of the prior distribution is equal to the REML estimate of the corresponding parameter obtained by AP. The degrees of freedom are chosen to minimize the weight of the prior distribution on the posterior inference. The Gibbs sampler algorithm of Section 3.3.2.3 is used to generate samples from the joint posterior distribution. We used a batching scheme so that the entire vector $\mathbf{b}_k$ was simulated at one Gibbs step. Batching is compared to a single-element scheme in Section 5.5.4.3. For this model and algorithm, the MPSR diagnostic indicated convergence after 19,000 iterations.

Posterior means of the location parameters and variance components for model M42 are listed in the second column of Table 5.4. Figure 5.3 shows histograms of the posterior distributions of the variance components, where it is evident that posterior modes of variance components are similar to REML estimates of variance components, especially for σ_ϵ^2 and the individual variance at a certain day. The table and figure indicate that REML-BLUP estimates and posterior means are reasonably similar. The 2.5%, 50% and 90% quantiles of the posterior distribution for each parameter are listed in Table 5.5. The difference between males and females are illustrated by their population average curves in Figure 5.4. Figure 5.4 also indicates great similarity between these two inference approaches in the male and female population curves.

Table 5.3 The priors and density function for random regression models used for fitting the pig weight gain data.

Model	parameter	prior
M42	σ_e^2	$I\chi^2(\nu_e = 4, \sigma_0^2 = 0.5^{(1)})$
	$\mathbf{G}$	$IW(\nu_g = 5, \mathbf{G}_0^{-1})^{(2)}$
	$\mathbf{b}$	constant
	$\mathbf{u}\|\mathbf{G}$	$N(\mathbf{0}, \mathbf{G})$
	$\mathbf{y}\|\mathbf{b}, \mathbf{u}, \sigma_e^2$	$N(\mathbf{Xb} + \mathbf{Zu}, \sigma_e^2\mathbf{I})$
M42R	σ_e^2	$I\chi^2(\nu_e = 4, \sigma_0^2)$
	$\mathbf{G}$	$IW(\nu_g = 5, \mathbf{G}_0^{-1})$
	$\mathbf{b}^+$	constant
	$\boldsymbol{\eta}\|\mathbf{G}$	$N(\tilde{\mathbf{b}}, \mathbf{G})$
	$\tilde{\mathbf{b}}$	constant
	$\mathbf{y}\|\mathbf{b}^+, \boldsymbol{\eta}, \sigma_e^2$	$N(\mathbf{Xb}^+ + \mathbf{Z}\boldsymbol{\eta}, \sigma_e^2\mathbf{I})$
M22R	σ_e^2	$I\chi^2(\nu_e = 4, \sigma_0^2)$
	$\mathbf{G}$	$IW(\nu_g = 5, \mathbf{G}_0^{-1})$
	$\boldsymbol{\eta}\|\mathbf{G}$	$N(\mathbf{b}, \mathbf{G})$
	$\mathbf{y}\|\boldsymbol{\eta}, \sigma_e^2$	$N(\mathbf{Z}\boldsymbol{\eta}, \sigma_e^2\mathbf{I})$

$I\chi^2$ stands for the inverse chi-square distribution,
and IW stands for the inverse Wishart distribution.
(1): in order to let mean=1.05 = $\hat{\sigma}^2_{e,REML}$.
(2): $\mathbf{G}_0$=$\hat{\mathbf{G}}_{REML}$ estimate/(5-3-1)

Since animal values estimated by model M42 includes heritable and non-heritable effects, it is not of interest to make inference about animal genetic performance at this point. Inferences for genetic values will be given in a subsequent section when the relationships among animals are incorporated.

5.5.3.2 Discussion

The purpose of this section is to discuss likelihood-based and Bayesian inferences. Comparisons among MCMC algorithms will be discussed later. First, we have verified the fact that when all random components follow a Gaussian distribution and a flat prior is assigned to the fixed effects, the posterior modes for model coefficients are the same as

Table 5.4 Comparison of the posterior means of parameters for a Bayesian analysis of models M42, M42L, M42RL with REML-BLUP estimates (AP).

Parameter	Model[1]			
	AP	**M42**	**M42L**	**M42RL**
b_{0_M}	0.224	0.1983	0.1980	0.1934
b_{1_M}	0.651	0.6540	0.6557	0.6542
b_{2_M}	0.0108	0.01072	0.01072	0.01071
b_{3_M}	-0.00014	-0.000139	-0.000139	-0.000139
b_{4_M}	5.53e-7	5.475e-7	5.473e-7	5.466e-7
b_{0_F}	0.302	0.3281	0.3259	0.3273
b_{1_F}	0.628	0.6257	0.6254	0.6252
b_{2_F}	0.0142	0.01425	0.01424	0.01424
b_{3_F}	-0.00018	-0.0001847	-0.0001846	0.0001845
b_{4_F}	7.29e-7	7.341e-7	7.340e-7	7.335e-7
G_{00}	0.977	0.9322	0.9332	0.9344
G_{11}	0.0149	0.01483	0.01484	0.01482
G_{22}	1.68e-6	1.663e-6	1.663e-6	1.659e-6
G_{01}	-0.0632	-0.06116	-0.06117	-0.06126
G_{02}	0.000432	0.000414	0.000413	0.0004124
G_{12}	-0.000117	-0.0001163	-0.0001163	0.0001160
σ_e^2	1.0546	1.0574	1.0574	1.0621
$\gamma^{(2)}$		19.000	8.000	400

AP: REML-BLUP estimates (Andersen and Petersen,1996)
M42: quartic-quadratic random regression model
R: reparameterization based on hierarchical centering
L: reparameterization on the Legendre transformation scale
(1): See Table 5.1 in Section 5.2 for more details about the models
(2): Convergence point

Table 5.5 Quantiles of the posterior distribution of the fixed effect parameters and variance components for models M42 and M42RL

	Posterior Quantiles					
	Model M42			Model M42RL		
Parameter	**2.5%**	**50%**	**97.5%**	**2.5%**	**50%**	**97.5%**
b_{0_M}	-0.1155	0.1957	0.4959	-0.1142	0.1937	0.5080
b_{1_M}	0.6158	0.6557	0.6948	0.6144	0.6542	0.6936
b_{2_M}	0.00948	0.01072	0.01194	0.00947	0.01071	0.01194
b_{3_M}	-1.57E-4	-1.39E-4	-1.22E-4	1.57E-4	-1.39E-4	-1.21E-4
b_{4_M}	4.643E-7	5.471E-7	6.305E-7	4.609E-7	5.4671E-7	6.311E-7
b_{0_F}	0.00648	0.3251	0.6469	.00749	0.3274	0.64507
b_{1_F}	0.5843	0.6254	0.6664	0.5843	0.6250	0.6669
b_{2_F}	0.01291	0.01424	0.01559	0.01292	0.01424	0.01557
b_{3_F}	-2.04E-4	-1.84E-4	-1.65E-4	-2.04E-4	-1.84E-4	-1.65E-4
b_{4_F}	6.392E-7	7.333E-7	8.292E-7	6.397E-7	7.332E-7	8.273
G_{00}	0.6728	0.9229	1.2474	0.6747	0.9244	1.2526
G_{11}	0.0118	0.0147	0.0185	0.0118	0.0147	0.0185
G_{22}	1.31E-6	1.65E-6	2.11E-6	1.30E-6	1.64E-6	2.10E-6
G_{01}	-0.088	-0.0603	-0.0382	-0.0884	-0.0606	-0.0383
G_{02}	1.76E-4	4.06E-4	6.88E-4	1.77E-4	4.06E-4	6.85E-4
G_{12}	-1.518E-4	-1.152E-4	-8.699E-4	-1.514E-4	-1.148E-4	-8.677E-4
σ_ϵ^2	1.0099	1.0568	1.1070	1.0140	1.0616	1.1117

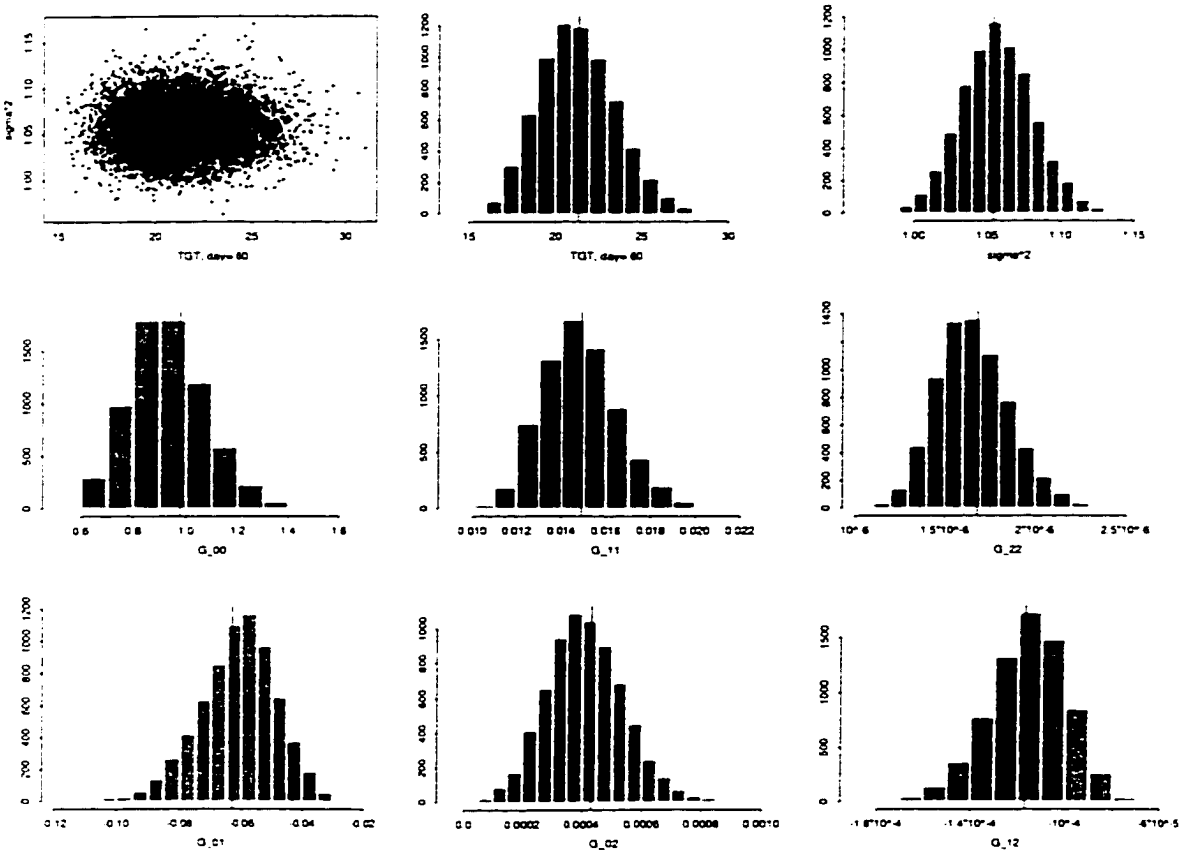

Figure 5.3 Histograms of parameter distributions of variance components and the animal variance on day 60 for model M42 (with REML estimates indicated by the vertical line). The top left figure is a scatter plot of σ_e^2 against animal variance on day 60

BLUP estimates obtained by solving Henderson's mixed model equations (MME) given the REML estimates for random components as mentioned in Chapter 3 (Table 5.4). For the fixed effects, consistency of posterior means and REML-BLUP results are shown in Table 5.4 and Figure 5.4. The variance for each model coefficient calculated from the posterior draws is approximately equal to the corresponding variance obtained from the inverse of the coefficient matrix in Henderson's MME (Harville, 1977; Robinson, 1991).

For comparing the two inference approaches, we can exercise the estimated animal

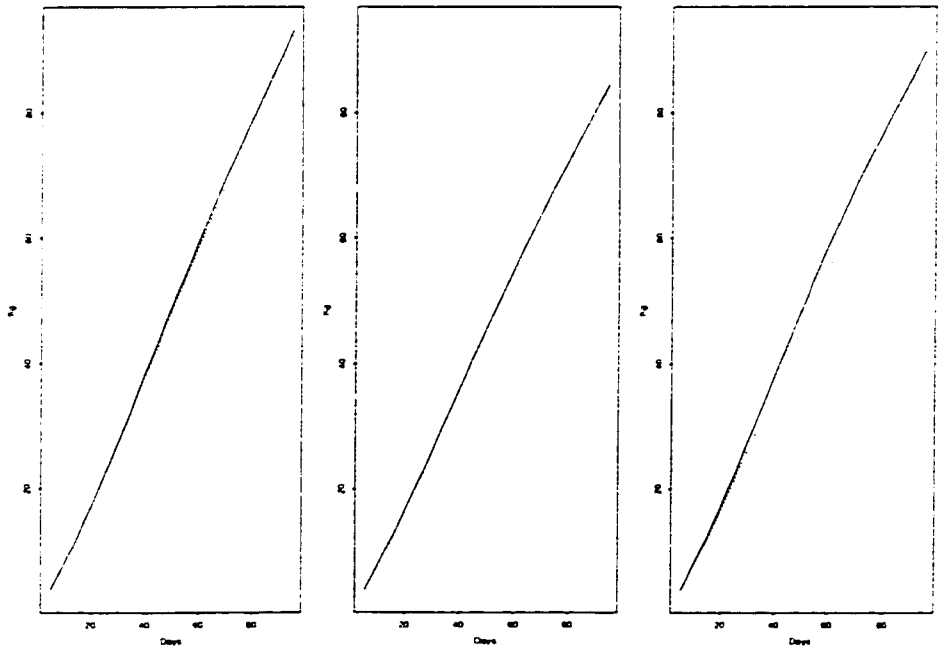

Figure 5.4 The population average curves for males and females obtained by REML-BLUP or Bayesian approaches to model M42. (left: for females, REML-BLUP (solid line), Bayesian (dashed line): middle: for males, REML-BLUP (solid line), Bayesian (dashed line): right: males (dashed line) females (solid line) by Bayesian approach. The two approaches yield estimates for males that are too similar to be distinguished here.

effects to see the consistency of the animal rankings produced by the two methods. The ranking obtained with REML-BLUP estimates is exactly the same as that obtained by ranking the posterior mean of animal effects (for any day on which ranking was attempted). This can also be explained by the similarity between the animal value estimates for the two approaches. The correlation between these two sets of estimates evaluated either on day 50, 75, or 100, is 0.9999.

It is impossible to calculate quantiles for REML-BLUP estimates without large sample theory. While any quantile for a quantity of interest can directly be obtained by

Bayesian approach. In terms of computing, the Bayesian approach reduces the need to calculate the inverse of large matrices. For example, the REML-BLUP approach requires the inverse of the large 580×580 coefficient matrix in Henderson's MME (Section 3.2) for this pig weight data set. By contrast, the Bayesian approach only requires the inverse of 3×3 or 5×5 matrices in some conditional distributions.

5.5.4 Comparing MCMC algorithms – independent animal model

The results in the previous section were obtained by fitting model M42 with the Gibbs sampling algorithm (plus batching). This took 19,000 iterations to converge. It is natural to consider if a more efficient algorithm is possible for fitting random polynomial regression models for large data sets. This section is focused on comparisons among various MCMC algorithms, which are related to orthogonality, hierarchical centering and batching strategies. The efficiency of MCMC algorithms is measured by the convergence point γ (see Section 5.2.3).

5.5.4.1 Orthogonal polynomials

The random polynomial regression model on the Legendre scale is denoted by MpqL. To transform model M42 to Model M42L, quadratic and quartic Legendre polynomials are used to adjust the time scale to be on the Legendre scale (see Section 5.4). So the block matrices in $\mathbf{X}_i$ and in $\mathbf{Z}_i$ are $\mathbf{T}_i\mathbf{L}_4^{-1}\mathbf{\Lambda}_4$ and $\mathbf{T}_i\mathbf{L}_2^{-1}\mathbf{\Lambda}_2$, respectively. Refer to Section 5.4 for the definition of $\mathbf{\Lambda}_4$; $\mathbf{\Lambda}_2$ is the top left 3×3 matrix of $\mathbf{\Lambda}_4$. Posterior autocorrelations among the fixed effects and the unique random components in $\mathbf{G}$ are reduced when this orthogonal transformation is applied to either model M42 or model M42R. Table 5.6 illustrates this, with the second line for model M42 and third line for model M42R.

Table 5.7 shows the benefit of using Legendre polynomials by comparing the convergence point for model M42L (γ=8,000) with that for model M42 (γ=19,000) (discussed

in 5.5.4.3). Figure 5.5 also shows the effect of orthogonality on the value of $\sqrt{\hat{R}^p}$, which remains smaller than 1.2 after 8,000 iterations.

Orthogonality is in particular beneficial when the fixed effects are drawn by the single-element scheme (γ >80,000 for the regular scale, γ=42,000 for the Legendre scale). Moreover, orthogonality makes an enormous difference when Legendre polynomials are adopted for the hierarchically centered model M42R (discussed further in 5.5.4.2), with γ reducing to 400 from >80,000 .

Table 5.6 Autocorrelation of parameters and convergence rate for models M42 and M42R

Covariate Scale	Model	Autocorrelation			γ	$s^{(3)}$
		σ_e^2	G(i,j)$^{(1)}$	Fixed Effects		
Regular	M42(single b)	0.133	0.190 ~ 0.558	0.975 ~ 0.999	>80,000	1,000
	M42(Batch **b**)	0.135	0.188 ~ 0.569	0.185 ~ 0.431	19,000	1,000
	M42R	0.134	0.191 ~ 0.559	0.580 ~ 0.998	>80,000	1,000
Legendre	M42(single b)	0.135	0.038 ~ 0.295	0.482 ~ 0.996	42,000	1,000
	M42(Batch **b**)	0.136	0.040 ~ 0.292	0.049 ~ 0.365	8,000	1,000
	M42R	0.145	0.044 ~ 0.279	0.033 ~ 0.783	400	200

(1) Range across six unique entries in the covariance matrix **G** for $\mathbf{u}_i$ or $\boldsymbol{\eta}_i$.
(2) γ: convergence point
(3) s: the interval length for sequential convergence diagnosis

5.5.4.2 Hierarchical centering

Gelfand et al.'s (1995a) hierarchical centering algorithm is adopted as follows. Since the degree of the polynomial for random effects **u** is less than that for fixed effects **b** for model M42, a partial reparameterized model can be applied by centering the random effects at the means of the corresponding fixed effects instead of at zeros. This partial reparameterized model is denoted by M42R and the model for a single observation is of

Table 5.7 Convergence rate[1] of 29 selected parameters[2] for models M42 and M42R

Scale of Polynomials	Sampling scheme for b		
	Single-element b_i	Batching b	
	M42s[3]	M42	M42R
Regular	>80.000	19.000	>80.000
Legendre	42.000	8.000	400

(1): indicated by convergence point γ (see Table 5.2)
(2): 29 parameters are 10 fixed effects, 7 random components, the average of random effects for each gender, and the random effects of the first 2 animals.
(3): M42s denotes model M42 with single-element sampling scheme.

the form

$$\begin{aligned} y_{ij} &= b^3_{3k(i)t_{ij}} + b_{4k(i)}t^4_{ij} + \eta_{0_i} + \eta_{1_i}t_{ij} + \eta_{2_i}t^2_{ij} + \epsilon_{ij} \\ &= \mathbf{t}^T_{4.2_{ij}}\mathbf{b}^+_{k(i)} + \mathbf{t}^T_{2_{ij}}\boldsymbol{\eta}_i + \epsilon_{ij}. \end{aligned} \tag{5.11}$$

where $\mathbf{t}_{4.2_{ij}} = (t^3_{ij}\ t^4_{ij})^T$, $\mathbf{t}_{2_{ij}} = (1, t_{ij}\ t^2_{ij}\ )^T$, $\boldsymbol{\eta}_i = (\eta_{0_i}, \eta_{1_i}, \eta_{2_i})^T = (b_{0k(i)}\ b_{1k(i)}\ b_{2k(i)})^T + (u_{0_i}\ u_{1_i}\ u_{2_i})^T = \tilde{\mathbf{b}}_{k(i)} + \mathbf{u}_i$ ($k(i)$=M or F for animal i's gender), $\mathbf{b}^+_{k(i)} = (b_{3k(i)}\ b_{k(i)4})^T$, $\mathbf{b}_{k(i)} = (\tilde{\mathbf{b}}^T_{k(i)}\ \mathbf{b}^{+T}_{k(i)})^T$, $\boldsymbol{\eta}_i \sim N(\tilde{\mathbf{b}}_{k(i)}, \mathbf{G})$, $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G})$, u_{j_i} are the same as those in (5.9), and $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$. The whole observation vector is expressed as (5.5).

$$\mathbf{y} = \mathbf{X}^+\mathbf{b}^+ + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{5.12}$$

where the rows in $\mathbf{Z}_i$ are $\mathbf{t}^T_{2_{ij}}$, and the rows in $\mathbf{X}^+_i$ are $\mathbf{t}^T_{4.2_{ij}}$ in the columns for the corresponding gender and zero elsewhere.

When fitting the random effects with the same degree of polynomial (i.e., allowing every coefficient to vary animal by animal), the complete centered model MqqR is $\mathbf{y} = \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}$, where the rows in $\mathbf{Z}_i$ are $\mathbf{t}_{q_{ij}} = (1\ t_{ij}\ \ldots\ t^q_{ij})$ in the columns for the corresponding gender.

In Table 5.7 and Figure 5.6, we find that hierarchical centering doesn't show a benefit in convergence rate for model M42R (in fact, there was no convergence before 80.000

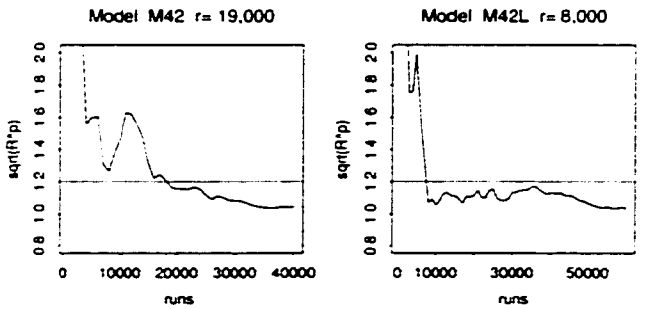

Figure 5.5 $\sqrt{\hat{R}^p}$ (estimate of MPSR) plots for model M42 and M42L

iterations), but hierarchical centering does make a big difference when incorporates with Legendre polynomials (γ= 8.000 for model M42L versus γ=400 for model M42RL). We will discuss why this combination enhances convergence rate later. Table 5.8 shows the benefit of hierarchical centering when fitting the completely centered model M22 or M44. The increase in convergence rate is dramatic.

5.5.4.3 Batching

There are two sampling schemes for drawing the fixed effects: batching $\mathbf{b}_{k(i)}$, ($k(i) = For.M$) and single-element b_i. Table 5.7 shows that the single-element scheme does not show convergence before 80.000 iterations, while the convergence point γ is 19.000 iterations when fitting M42 with a batching scheme (Table 5.7, Figure 5.5). Therefore, the posterior samples obtained from the batching scheme can be used to compare with REML-BLUP results. The dimension for $\mathbf{u}$ is often large, so that batching subvector

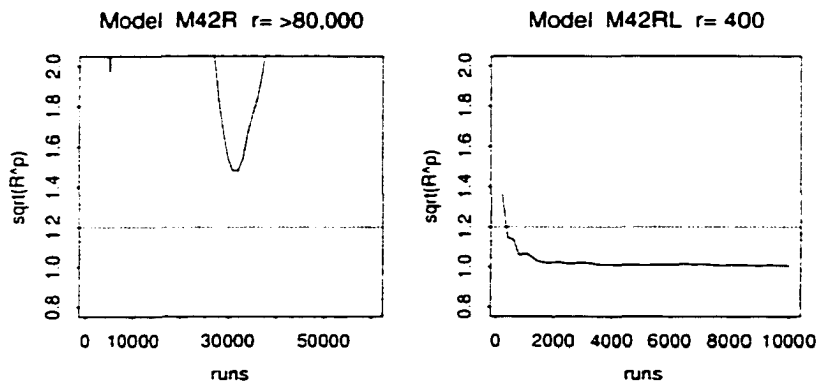

Figure 5.6 $\sqrt{\hat{R}^p}$ (estimate of MPSR) plots for models M42R and M42RL.

$\mathbf{u}_i$ is sensible than batching whole $\mathbf{u}$, especially when animals are unrelated. Thus we always do this. It is not clear whether it is beneficial.

Posterior correlations between the fixed effects for model M42 are very large (with values from 0.975 to 0.999, see Table 5.6). Batching fixed effects results in a great reduction in the posterior autocorrelation among the fixed effects (posterior correlations then varying from 0.185 to 0.431). When the Legendre scale is adopted, batching still yields lower posterior correlations among the fixed effects than the single-element sampling scheme (range= $0.049 \sim 0.365$ vs $0.482 \sim 0.996$).

The batching scheme for drawing fixed effects results in better convergence rate than the single-element scheme (γ=19.000 vs >80.000) for model M42 on the regular scale. When the Legendre scale is adopted, the batching scheme for $\mathbf{b}$ is still better than the

Table 5.8 Convergence rate[1] of selected parameters[2] for models M44, M42R, M22 and M22R.

Scale of Polynomials	Model			
	M44	M44R	M22	M22R
Regular	>80.000	13.000	16.000	300

(1): indicated by convergence point γ (see Table 5.2)
(2): parameters are the fixed effects, random components, the average of random effects for each gender, and the random effects of the first 2 animals.

single-element scheme (γ=8.000 vs 42.000). This shows the benefit of drawing correlated parameters together (Gilks and Roberts, 1996).

5.5.4.4 Summary and discussion

Our results comparing convergence rates of different MCMC algorithms for fitting pig weight gain data with the random polynomial regression model M42 can be summarized as follows (see Table 5.7): (1) batching the fixed effect parameters improves convergence rate; (2) hierarchical centering doesn't improve the convergence rate for model M42 on the regular scale, but it has a great effect on models with the Legendre scale; (3) the use of orthogonal polynomials shows a benefit in convergence; (4) the greatest improvement in convergence rate is obtained when both hierarchical centering and orthogonality are applied to the random regression model; and (5) hierarchical centering is very efficient when the same degree of polynomial is fitted to the fixed and random effects.

For random polynomial regression models, the fixed effects are correlated. Drawing posterior samples by batching is equivalent to traveling by a multi-dimensional movement. This is efficient because it maintains the relationship among the elements of the parameter vector (Gilks and Roberts, 1996). Note that each step of the batching scheme is a little slower since a 5×5 matrix is inverted in model M42, but that seems not to

be a large burden. Thus we observe a benefit of batching fixed effects in our study.

Theoretically, batching and single-element schemes are expected to have the same convergence rate for orthogonal polynomial models for balanced data, since parameters will be independent of each other. The results from fitting M42L shows that batching fixed effects still improves convergence rate (γ=8.000 vs 42.000). The use of orthogonality for the unbalanced pig weight gain data improves the convergence rate to some extent when the fixed effects are drawn element by element (γ= >80.000 vs 42.000).

Autocorrelation is another indicator for convergence rate. In general, the autocorrelation is significantly reduced when orthogonal polynomials are used. We have checked several lag-autocorrelations for every set of posterior samples and observe that the autocorrelation of lag 20 is greatly reduced.

Reparameterization strategies are designed to reduce posterior correlations between parameters, such as hierarchical centering (Gelfand et al. 1995a) and the use of orthogonality (Gilks and Roberts, 1996). When the Gibbs sampler is adopted to explore the joint posterior distribution, the more independence among parameters, the faster travel throughout the parameter space will be. Therefore, batching or orthogonality reduces the correlation and thus improves convergence rate.

The reason that hierarchical centering (see Gelfand et al., 1995a) doesn't perform its benefit for the data analyzed here can be due to the small magnitude of σ_e^2 relative to $var(\mathbf{u}_i) = \mathbf{G}$ (Table 5.4). Neither is particularly large. Hierarchical centering shows its benefit when orthogonal polynomials are adopted because on the Legendre scale the variance matrix ($\mathbf{G}^*$) is much larger than σ_e^2 (on regular scale for both cases). Note that orthogonal transformation also reduces the posterior correlation among the fixed effect parameters and random effect parameters when the hierarchical centering algorithm is adopted.

5.5.5 Bayesian analysis with dependent animal model

The data set used here are from half-sib and full-sib families. Thus, it is necessary to incorporate genetic relationships among animals in the model to obtain accurate estimates of animal genetic parameters. Therefore, model M42A (see (5.4)in Section 5.2.2) with genetic and permanent environmental random effects should be used. Recall that the model is $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zp} + \boldsymbol{\epsilon}$.

5.5.5.1 Incorporating genetic relationships

Model M42A is fitted to pig weight gains when additive genetic relationships among animals (given in the $n \times n$ matrix $\mathbf{A}$) are incorporated. Because of the significant impact of hierarchical centering and orthogonality on the convergence rate for the model without animal relationship (model M42), a hierarchically centered model on the Legendre scale, M42RAL, is used to fit the pig weight gains when relationships among animals are taken into account. The fixed effects are drawn using the batching scheme.

Note that there are two random components besides the random residual. Therefore, there are two options with respect to hierarchical centering: either centering the genetic effect parameters or centering the permanent environmental effect parameters. The algorithms of centering genetic and permanent environmental effect parameters are denoted by algorithm ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{p}$) (with $\boldsymbol{\eta}_i = \tilde{\mathbf{b}}_K + \mathbf{a}_i$) and algorithm ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{a}$) (with $\boldsymbol{\eta}_i = \tilde{\mathbf{b}}_K + \mathbf{p}_i$), respectively. According to Gelfand and Carlin's study (1995), centering the set of random effects with the largest variance would yield the most efficient algorithm. This suggests that centering the permanent environmental effect parameters seems more sensible in our case, since the heritability for pig weight gain is around 0.2 to 0.35. However, Gelfand and Carlin's results are for iid random effects. Strong relationships between animals could influence the relative magnitude of the posterior genetic variance to the posterior environmental variance, which is the key quantity for assessing relative efficiency of cen-

tering (see Section 4.2.3). We have tried both algorithms with the pig weight gain data, but neither converge quickly (γ is at least 72.000). It is desirable to find a more efficient algorithm for fitting model M42RAL.

In the next section, we compare algorithms before presenting the results for pig weight gain data when genetic relationships are incorporated. Let $\mathbf{G}_a$ and $\mathbf{E}$ denote the variance matrices for the animal additive genetic effects $\mathbf{a}_i$ and the permanent environmental effects $\mathbf{p}_i$, respectively. We assume that $\mathbf{G}_a = r\,(\mathbf{G}_a + \mathbf{E})$ with r indicating the ratio of the genetic variance $\mathbf{G}_a$ to the total animal variance $\mathbf{G}_a + \mathbf{E}$.

5.5.5.2 Comparing MCMC algorithms – genetic/environmental effects model

As mentioned in Section 5.5.5.1, the convergence rate for fitting model M42RAL by Bayesian approach is slow when either centering genetic effect parameters or centering permanent environmental effect parameters is adopted. Therefore, it is of interest to explore the merits of the cycling algorithm introduced by Gelfand and Carlin (1995). Recall that the cycling algorithm implements algorithms ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{a}$) and ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{p}$) in sequence, as described in Section 4.2.3.

In order to understand whether the efficiency of the cycling algorithm would be affected by the relative ratio of genetic variance to permanent environmental variance, five data sets of 1,000 animals with known relationships between animals are simulated for model (5.4) by varying the ratio r from 0.1 to 0.9 (recall that $\mathbf{G}_a = r\,(\mathbf{G}_a + \mathbf{E})$). The parameter values used to generate the data sets are the REML-BLUP estimates obtained from pig weight gains (see Table 5.4). The true value of $\mathbf{G}_a$ for each simulated data is obtained by multiplying the ratio r with the overall $\mathbf{G}_{REML}$ (obtained by ignoring genetic relationships).

The three hierarchical centering algorithms, ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{a}$), ($\mathbf{b}\ \boldsymbol{\eta}\ \mathbf{p}$), and cycling, are applied to each of five simulated data sets. Convergence rates for these three algorithms

implemented for model M42RAL are listed in Table 5.9. Results show that the cycling algorithm works better or at least no worse than the algorithms (**b** η **a**) or (**b** η **p**), no matter what the ratio r is.

We found that even with 1,000 animals the choice of prior distributions for **G** and **E** had a large impact on the posterior distributions of **G** and **E**. For the results presented here the scale matrix in the inverse-Wishart prior distributions were chosen in the same proportion (r) as the values used to generate the data. Of course, this is not possible in general.

Table 5.9 Comparisons among three hierarchical centering algorithms for model M42RAL[1] in terms of the convergence rate[2] for simulated data sets with various r[3].

	Algorithm		
r	η=b+a b − η − p	η=b+p b − η − a	Cycling
0.1	62.000	>80.000	26.000
0.3	>80.000	63.000	40.000
0.5	56.000	63.000	40.000
0.7	31.000	73.000	30.000
0.9	38.000	76.000	31.000

(1): Refer to Table (5.1)
(2): Refer to Table (5.2)
(3): $\mathbf{G}_a = r(\mathbf{G}_a+\mathbf{E})$, in the simulated data with $\mathbf{G}_a+\mathbf{E} = \mathbf{G}_{REML}$.

5.5.5.3 Results for pig weight gains

Model M42RAL (which incorporates genetic relationships) is used to fit the pig weight data. Based on the simulation study in the previous section, the cycling algorithm is used to analyze the pig weight gain data. Since the value of r affects the estimation of $\mathbf{G}_a$ for small data sets (recall that there are only 190 animals in this data set). A range of values are assigned for the prior ratio for r, from 0.2 to 0.35 according to the accepted

range of heritability for pig weight gain. With five Markov chains, the convergence point is about 36,000 iterations. Posterior means and quantiles for location parameters and variance components are listed in Table 5.10. Posterior means of the fixed effects and residual variance are similar to those obtained by model M42 (in Table 5.4) no matter which value is used for the prior ratio r. However, the genetic variance and permanent environmental variance vary with the prior ratio r. This is a result of the small number of animals in the data set.

Since use of an incorrect value for r in choosing prior distribution scale matrices can result in incorrect inference, the impact of r on the ranking of animals is evaluated next. Animal ranking is evaluated for r varying from 0.1 to 0.45. The results indicatet that the top 10 animals for each gender based on posterior means of their genetic values are the same when r is between 0.2 to 0.35, but there are some differences in ranking when r is not within this range. The top 10 animals for each gender evaluated by genetic value are listed in Table 5.11 under the heading ID, when $r = 0.2 \sim 0.35$ is assumed. Note that for these models the ranking depends on the day we considered. Results are presented for days 50, 75, and 100. Table 5.11 indicates differences in ranking when M42 (ignoring relationship) or the nonlinear model (of Chapter 6) are used. The top 10 animals are understandablely different from the top 10 animals ranked when relationships among animals are ignored, since the ranking for the latter is based on the sum of genetic and permanent environmental effects.

It is also of interest to have a confidence region for the contribution of an individual's genetic value over time. The estimated genetic value of individual i on day t is calculated by $\hat{g}_i = \hat{a}_{0_i} + t\hat{a}_{1_i} + t^2\hat{a}_{2_i}$, where $\hat{a}_{j_i}; j = 0, 1, 2$, are posterior simulations of the genetic effects for animal i. Then the 95% confidence region for $\hat{g}$ of each animal can be obtained by calculating $\hat{g}$ from its posterior draws of $\mathbf{a}$ at each iteration and then finding the 2.5% quantile, median, and 97.5% quantile for $\hat{g}$ on day t. Confidence regions for the genetic values of the top 6 animals for each gender on day 75 are shown in Figure 5.7. It

Table 5.10 Posterior means of selected parameters for model M42RAL[1] fitted to pig weight gains with prior scale measures for $\mathbf{G}_a$ and $\mathbf{E}$ such that the ratio of genetic variance to animal variance is r, with r varying from .2 to .35.

	Prior ratio r			
Parameter	**0.2**	**0.25**	**0.3**	**0.35**
b_{0_M}	0.1912	0.1903	0.1900	0.1900
b_{1_M}	0.6535	0.6534	0.6533	0.6533
b_{2_M}	0.01072	0.01072	0.01072	0.01072
b_{3_M}	-0.000139	-0.000139	-0.000139	-0.000139
b_{4_M}	5.473E-7	5.472E-7	5.472E-7	5.471E-7
b_{0_F}	0.3238	0.3230	0.3221	0.3218
b_{1_F}	0.6242	0.6241	0.6240	0.6239
b_{2_F}	0.01427	0.01427	0.01427	0.01427
b_{3_F}	-0.000185	-0.000185	-0.000185	-0.000185
b_{4_F}	7.351E-7	7.352E-7	7.353E-7	7.354E-7
G_{00}	0.1373	0.1616	0.1857	0.2093
G_{11}	0.00212	0.00242	0.00270	0.00299
G_{22}	2.906E-7	3.290E-7	3.649E-7	3.998E-7
G_{01}	-0.00905	-0.01050	-0.01185	-0.01319
G_{02}	0.000082	0.000095	0.000107	0.000120
G_{12}	-0.000019	-0.000022	-0.000025	-0.000027
E_{00}	0.797	0.777	0.756	0.736
E_{11}	0.01291	0.01266	0.01244	0.01223
E_{22}	1.388E-6	1.358E-6	1.330E-6	1.303E-6
E_{01}	-0.0522	-0.0510	-0.0499	-0.0488
E_{02}	0.000328	0.000317	0.000307	0.000297
E_{12}	-0.000098	-0.000096	-0.000094	-0.000092
σ_e^2	1.0592	1.0592	1.0593	1.0594
$\gamma^{(2)}$	35.000	36.000	36.000	36.000

(1): Refer to Table (5.1)

(2): An indicator for convergence rate, referring to Table (5.2)

Table 5.11 Ranks of the top 10 animals for each gender in terms of the posterior means of genetic values or animal values estimated by models M42RAL, M42 and NLM[1] on day 50, 75, and 100

Ranked by	Day 50			Day 75			Day 100		
M42RAL	ID[2]	M42	NLM	ID	M42	NLM	ID	M42	NLM
Male 1	104	3[3]	3[3]	77	2	2	29	2	12
2	152	4	4	29	3	7	11	1	7
3	77	2	2	59	1	1	176	9	
4	59	1	1	104	7	5	77	4	3
5	111	5	5	11	5		59	3	1
6	140	6	7	152			5	6	
7	108	7	6	108		4	61	10	
8	142	8	8	176			164		5
9	115			164			57	8	4
10	29			5			108	7	6
Female 1	151	4	10	151	4	10	56	1	1
2	185	2	2	56	3	3	8	4	
3	103			185	2	2	185		6
4	78	4	3	78			151	6	
5	141	3	4	8	10		94		
6	183			45		9	45	2	2
7	116	10		103			129		
8	54			54			90	5	
9	155			155			14		
10	99			116	5	8	30		

(1): NLM stands for nonlinear model discussed in Chapter 6.
(2): Animal ID (3): Ranked by the corresponding model

illustrates that some animals (e.g., females with id= 8, 56) grow slower in terms of genetic value than the population genetic average in the beginning but faster at a later stage. Other animals (e.g., id= 104, 151 in the first row) grow faster than population average over the entire experimental period.

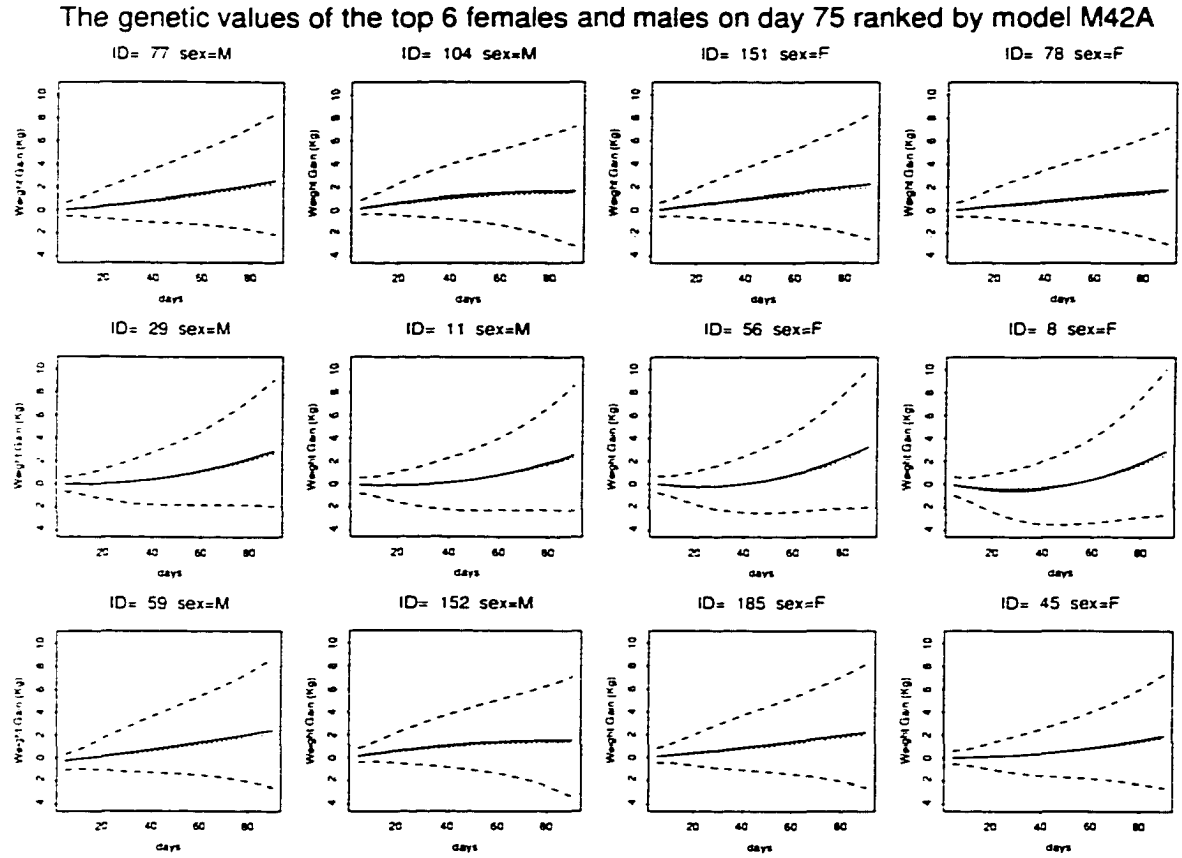

Figure 5.7 The 95% posterior region for the genetic value for the top 6 animals of each gender (Solid line is the posterior mean genetic value, dotted line is the posterior median genetic value).

Heritability (h^2) is one of the most important breeding characteristics. Heritability is defined as the ratio of genetic variance to phenotypic variance and therefore is a function

of population parameters. For model M42A, h^2 on day t can be estimated by

$$h^2 = \frac{(1\ t\ t^2)\mathbf{G}_a(1\ t\ t^2)^T}{(1\ t\ t^2)(\mathbf{G}_a + \mathbf{E})(1\ t\ t^2)^T + \sigma_\epsilon^2}.$$

where $\mathbf{G}_a$, $\mathbf{E}$, and σ_ϵ^2 are posterior samples. The heritability h^2 can be computed for each iteration of the posterior samples, which allows us to characterize the distribution of h^2. The mean, standard error, and some quantiles for h^2 on day 50, 75 or 100 are given in Table 5.12. The posterior distribution for h^2, the genetic variance, and the phenotypic variance on day 75 are shown in Figure 5.8. This figure indicates that the distribution of h^2 and genetic variance are not symmetric, as is evident from the significant differences between mean and median.

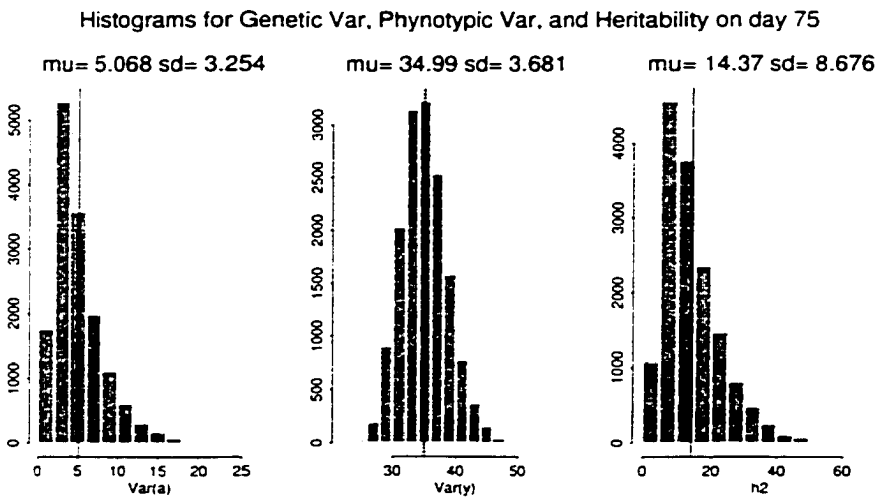

Figure 5.8 Histograms showing posterior distributions of the genetic variance, the phenotypic variance, and heritability on day 75 obtained by fitting model M42RAL. (median: dotted vertical line, mean: solid vertical line).

Table 5.12 Characteristics of the distribution of the heritability of pig weight gain summarized by posterior samples on day 50, 75, and 100, with prior r=0.25.

	Heritability h^2							
Day	Mean	2.5%[1]	Medium	97.5%[1]	se(h^2)	Correlation[2]		
50	14.05	3.36	11.02	43.44	10.18		0.7919	0.3537
75	14.38	3.81	12.21	36.60	8.68	0.8971		0.7957
100	15.93	4.19	13.91	39.09	9.13	0.5479	0.8607	

(1): quantiles
(2): upper triangle: correlation between h^2's on three selected days.
: lower triangle: correlation between genetic values.

5.5.6 Summary and discussion

The Gibbs sampler is used to provide inferences from the joint posterior distribution of polynomial regression models. We consider orthogonal transformation, hierarchical centering, and batching in order to find efficient MCMC algorithms when the Bayesian approach to random polynomial regression models is adopted for longitudinal data. The study results are summarized as follows. (1) orthogonality (i.e., the use of orthogonal polynomials) plays a significant role in reducing correlation among parameters and therefore yields a large improvement in convergence rate of MCMC methods, no matter whether it is implemented alone or with other algorithms like hierarchical centering; (2) batching the fixed effect parameters for subpopulations is more efficient than drawing parameters element by element; (3) hierarchical centering (Gelfand et al. 1995a) may not be helpful in convergence rate when random polynomial regression models are fitted with different degree polynomials for the fixed and random effects. When fitting the same degree polynomial to the fixed and random effects, hierarchical centering can show its benefits in convergence rate; (4) it is especially helpful to use orthogonal polynomials along with hierarchical centering.

There are two random factors in the random regression models when the genetic

additive relationship between animals is incorporated to address genetic and permanent environmental effects. Convergence is slow when orthogonal hierarchical centering is applied to one of the two random factors. The cycling algorithm (Gelfand and Carlin, 1995) yields the best performance.

When flat prior distributions are assigned to the fixed effect parameters and the random effect parameters follow Gaussian distributions, there is a lot of similarity between likelihood-based and Bayesian inference (Harville, 1977; Robinson, 1991). The advantages of the Bayesian approach appear when drawing inferences for quantities, which are complicated functions of model parameters such as heritability, ranking, genetic values, etc., because posterior draws are available for exploring the characteristics of the posterior distributions of these quantities of interest. In many cases the distributions of such quantities do not have normal distribution, in which case inferences drawn from the Bayesian approach are likely different from those that might be drawn from the REML-BLUP approach, which relies on large sample theory.

CHAPTER 6 NONLINEAR MIXED MODELS FOR LONGITUDINAL DATA

6.1 Introduction

In Chapter 5 we focused on the Bayesian approach to data analysis using random polynomial regression models and the factors associated with efficient Markov chain Monte Carlo (MCMC) computation for those models. In this chapter we study the factors associated with efficiency of MCMC in the context of nonlinear mixed (NLM) models.

The random polynomial regression models describe nonlinear growth patterns but are linear in the model parameters. NLM models are often used to describe variation in biological characters with longitudinal data, e.g., growth and lactation traits. (Walkfield et al., 1994; Rodriguez-zas et al., 1997). The presence of a nonlinear function of the parameters in the likelihood means that different MCMC algorithms are required. Specifically, Gibbs sampling is no longer easy to implement because some of the conditional distributions are not known (Chib and Greenberg, 1995). Thus some Metropolis steps must be integrated within the Gibbs sampler. As a result, the factors affecting the efficiency of MCMC are different for NLM models than for linear models.

The layout for this chapter is as follows. Nonlinear mixed models are introduced in the remainder of this section. Bayesian model specification for NLM is provided in Section 6.2. A number of Metropolis-Hastings (M-H) algorithms for drawing posterior samples are described in Section 6.3. In Section 6.4 the efficiency of the various algo-

rithms are compared using a simulation study. The best of the algorithms are applied to pig weight gains in the final section to draw inferences about the quantities of interest.

6.1.1 Nonlinear functions

A nonlinear function is a function in which at least one of the parameters appears nonlinearly. Another characterization is that, in a nonlinear function, at least one of the derivatives of the response variable with respect to the parameters, is a function of at least one of those parameters (Ratkowsky, 1990). Note that the polynomial models of Chapter 5 are linear models in the sense described here.

Nonlinear functions have been proposed for modeling growth of a physical trait in terms of a small number of biologically interpretable parameters, $\boldsymbol{\theta}$. Ratkowsky (1990) reviews a set of nonlinear functions ranging from one-parameter models to four-parameter models to describe the relationship between response and time. The functions include some subset of the following parameters: an upper asymptote, a growth rate, an inflection point, and a lower asymptote. If growth is expected to follow a sigmoidal shape with lower asymptote zero, a finite upper asymptote, and an asymmetry about its inflection point, then the three parameter Gompertz function can be used to build the relationship between the responses and the explanatory variables. The Gompertz function is expressed as,

$$f(\boldsymbol{\theta}, t) = \eta \, exp\left\{-exp\left(-\beta\left(\frac{t}{\kappa} - 1\right)\right)\right\}, \tag{6.1}$$

where t is time, $\boldsymbol{\theta} = (\eta \ \beta \ \kappa)^T$ is the parameter vector with η denoting the upper asymptotic value of the response, β the growth rate, and κ the inflection point of the growth curve. Figure 6.1 presents several illustrative Gompertz curves. In this chapter, the Gompertz function is used to illustrate the fitting of nonlinear models via the Bayesian approach and to address issues concerning optimal algorithms for the Bayesian approach.

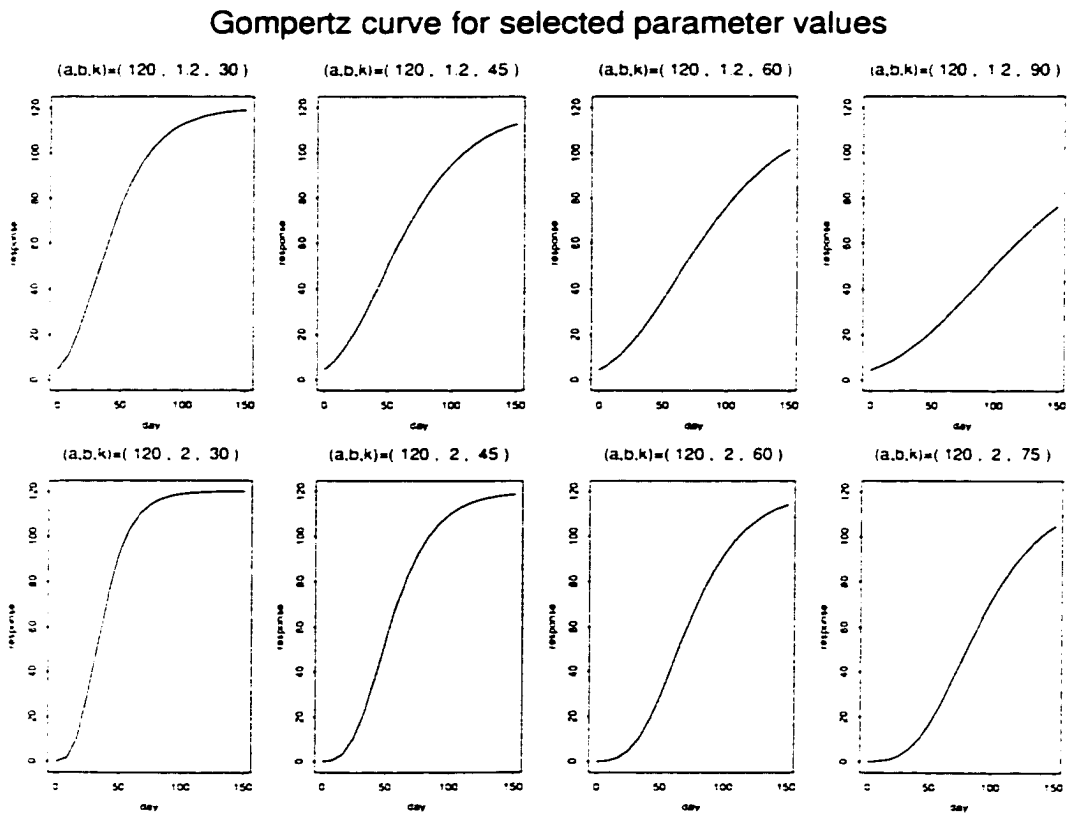

Figure 6.1 The Gompertz curve for selected parameter values.

6.1.2 Nonlinear mixed models

Nonlinear mixed (NLM) models are a natural extension of the linear mixed models of the previous chapter. The parameters of the nonlinear model for individuals are assumed to be random samples from a population. Let y_{ij} and t_{ij} be the j^{th} observation and the measuring time on individual i ($i = 1.2.\ldots.n; j = 1.2.\ldots.r_i$). Individuals are assumed to belong to subpopulations (e.g., males and females). Within each subpopulation, the parameter vectors for the nonlinear model are assumed to vary randomly about the subpopulation average. Then the NLM models for observation y_{ij}, belonging to

subpopulation $k(i)$ $(k(i) = 1, 2, \ldots, m)$, is written as

$$y_{ij} = f(\boldsymbol{\theta}_i, t_{ij}) + \epsilon_{ij}, \qquad \boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_{0k(i)}, \mathbf{G}). \tag{6.2}$$

where $f(.)$ is a nonlinear function with functional parameter vector $\boldsymbol{\theta}_i$, where $\boldsymbol{\theta}_{0k(i)}$ is the parameter vector for subpopulation $k(i)$, and $\mathbf{G}$ is the covariance matrix for the random effects $\boldsymbol{\theta}_i$.

Note that the nonlinear mixed model is specified here with hierarchically centered parameters. The alternative specification in which we write $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{0k(i)} + u_i$ is less practical in nonlinear models because subpopulation parameters $\boldsymbol{\theta}_{0k(i)}$ and individual parameters u_i (from $N(\mathbf{0}, \mathbf{G})$) would both be embedded in the nonlinear function $f(.)$.

6.1.3 Issues in implementing Bayesian methods for nonlinear models

The presence of a nonlinear function in the joint posterior distribution typically means that we can not directly sample from each of the conditional distributions required in Gibbs sampling (Gelman and Gelman, 1984). Instead, the Metropolis-Hastings (M-H) algorithm (Metropolis, 1953; Hastings, 1970) is required. The M-H algorithm can be used for one or more steps within a Gibbs sampler or a single M-H step may be used to replace the Gibbs sampler. A jumping distribution is required to implement the M-H algorithm (Chib and Greenberg, 1995). Since the jumping distribution determines the potential points to visit in the parameter space, issues associated with the choice of jumping distribution can have a large effect on the efficiency of MCMC. One measure of the efficiency of an MCMC algorithm is the number of iterations required until convergence. That is, until the generated samples provide an adequate approximation to the posterior distribution. If a jumping distribution does not jump far enough or to all regions supported by the posterior distribution, then the simulated sequence of draws would stay in a small region and not represent the posterior distribution.

There are many options for the choice of jumping distribution. A common approach is a random walk MCMC. In that case the jumping distribution does not formally depend on the likelihood function or on the posterior distribution. Instead, the jumping distribution is a specified distribution, usually symmetrical, centered at the current value of the parameter (or vector of parameters). The scale of the jumping distribution is adjusted to achieve the optimal acceptance rate, a rate between .23 and .45 according to results of Gelman et al. (1996). An alternative approach takes the jumping distribution as an approximate of the target distribution (see Section 4.4.1). In the context of nonlinear models, this can be done by constructing a linear approximation (Lindstrom and Bates, 1990; Wolfinger and Lin, 1997) to the nonlinear function in the likelihood (posterior distribution). Then a jumping distribution can be chosen to be centered at the mode of the approximation to the posterior distribution. The variance of the jumping distribution can be taken as the Hessian matrix or as a function of only the first derivative vector. These issues are described more fully in Section 6.3.

Other factors affecting the efficiency of M-H implementation are the organization of parameters into subvectors or batches and the choice of starting values. These are also described more fully in Section 6.3. Section 6.2 describes the Bayesian model in more detail. Then, Section 6.3 describes the algorithms that are compared and Section 6.4 provides the comparison results. Lastly, we compare the efficiency of different MCMC algorithms for carrying out a Bayesian analysis of simulated data sets and the data set of pig weight gains introduced in Chapter 5.

6.2 Bayesian approach to nonlinear mixed models

In this section, a Bayesian model for a growth trait, based on the nonlinear Gompertz family (see (6.1)) of growth curves is described. This model is the basis for the study of different MCMC algorithms.

For convenience, notation used in this section is defined in advance before introducing the model.

Notation

n= total number of animals

r_i= total number of repeated measurements on animal i: $i = 1, 2, \ldots, n$

n_k= total number of animals in subpopulation k

$N = \sum_i r_i$= the total number of records in the data set

$k(i) = 1, 2, \ldots, m$ = the indicator of the subpopulation which individual i belongs to.

$\boldsymbol{\theta}_i = (\eta_i \ \beta_i \ \kappa_i)$= the Gompertz function parameters for animal i.

y_{ij}= the observation taken on individual i at time t_{ij} $(j = 1, \ldots, r_i)$

$\mathbf{y}_i$= vector of observations taken on animal i

$f(\boldsymbol{\theta}_i, t_{ij})$= the expected value of y_{ij} for animal i at time t_{ij}, where $f(.)$ is the Gompertz function with parameter vector $\boldsymbol{\theta}_i$.

$\mathbf{y} = (\mathbf{y}_1 \ \mathbf{y}_2 \ \ldots \ \mathbf{y}_n)$ = the entire observed data vector

$$\mathbf{G} = var(\boldsymbol{\theta}_i) = \begin{pmatrix} G_{\eta\eta} & G_{\eta\beta} & G_{\eta\kappa} \\ G_{\eta\beta} & G_{\beta\beta} & G_{\beta\kappa} \\ G_{\eta\kappa} & G_{\beta\kappa} & G_{\kappa\kappa} \end{pmatrix}, \text{ and } \mathbf{G}^{-1} = \begin{pmatrix} G^{\eta\eta} & G^{\eta\beta} & G^{\eta\kappa} \\ G^{\eta\beta} & G^{\beta\beta} & G^{\beta\kappa} \\ G^{\eta\kappa} & G^{\beta\kappa} & G^{\kappa\kappa} \end{pmatrix}$$

6.2.1 Model specification

Suppose the progress of a growth trait over time can be well approximated by the Gompertz function (6.1). Then the data for individual i can be modeled using a nonlinear mixed model (6.2), $y_{ij} = f(\boldsymbol{\theta}_i, t_{ij}) + \epsilon_{ij}$, $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, r_i)$, where parameter vector $\boldsymbol{\theta}_i = (\eta_i \ \beta_i \ \kappa_i)^T \sim N(\boldsymbol{\theta}_{0k(i)}, \mathbf{G})$, with η the asymptotic value, β the growth

rate, and κ the inflection point, $\boldsymbol{\theta}_{0k(i)}$ the mean vector for the subpopulation, to which individual i belongs, and $\epsilon \sim N(0, \sigma_e^2)$. Under the assumptions of independence between animals and normal distributions for the random residuals ϵ, the likelihood of the entire observed data vector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ is

$$\begin{aligned} L(\boldsymbol{\theta}, \sigma_e^2 \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma_e^2) \\ &= \Pi_{i=1}^{n} \Pi_{j=1}^{r_i} p(y_{ij} \mid \boldsymbol{\theta}_i, \sigma_e^2) = \Pi_{i=1}^{n} \Pi_{j=1}^{r_i} N(y_{ij} \mid f(\boldsymbol{\theta}_i, t_{ij}), \sigma_e^2) \\ &\propto (2\pi\, \sigma_e^2)^{-N/2} exp\left\{ \frac{-1}{2\sigma_e^2} \sum_{i}^{n} \sum_{j=1}^{r_i} (y_{ij} - f(\boldsymbol{\theta}_i, t_{ij}))^2 \right\}. \end{aligned}$$

The individual parameter vectors $\boldsymbol{\theta}_i$ are assumed to be independent and drawn from population distribution $N(\boldsymbol{\theta}_{0k(i)}, \mathbf{G})$ $(k(i) = 1, 2, \ldots, m)$. The prior distributions for the remaining parameters in the NLM model are set up as follows.

- $\sigma_e^2 \sim I\chi^2(\nu_e, \sigma_0^2)$
- $p(\boldsymbol{\theta}_0) = p(\boldsymbol{\theta}_{0k}; k = 1, \ldots, m) \propto$ constant
- $\mathbf{G} = var(\boldsymbol{\theta}_i) \sim IW(\nu_g, \mathbf{G}_0)$

where $I\chi^2$ and IW indicate the inverse chi-square distribution and the inverse Wishart distribution (refer for their density functions to Section 5.2.1). The degrees of freedom parameters ν_e and ν_g are selected to lower the impact of the prior distribution on the inference. For example, ν_e=4.0 yields a prior with infinite variance.

The joint posterior distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \mathbf{G}, \sigma_e^2 \mid \mathbf{y}) &= p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma_e^2) p(\sigma_e^2 \mid \sigma_0^2) p(\boldsymbol{\theta} \mid \boldsymbol{\theta}_0, \mathbf{G}) p(\boldsymbol{\theta}_0) p(\mathbf{G} \mid \mathbf{G}_0)) \\ &= p(\sigma_e^2 \mid \sigma_0^2) p(\mathbf{G} \mid \mathbf{G}_0) \Pi_{k=1}^{m} p(\boldsymbol{\theta}_{0k}) \Pi_{i=1}^{n} \left\{ p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{0k(i)}, \mathbf{G}) p(\mathbf{y}_i \mid \boldsymbol{\theta}_i, \sigma_e^2) \right\}. \end{aligned}$$

6.2.2 Full conditional posterior distributions

The full conditional posterior distributions of parameters are as follows. Refer to Section (6.2) for the notation used here.

- $\sigma_e^2 \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\theta}_0, \mathbf{G} \sim I\chi^2\left(\nu_e + N, \left(\nu_e\sigma_0^2 + \sum_i \sum_j (y_{ij} - f(\boldsymbol{\theta}_i, t_{ij}))^2\right)/(N + \nu_e)\right)$

- $\mathbf{G} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\theta}_0, \sigma_e^2 \sim IW\left(\nu_g + n, (\mathbf{G}_0 + \mathbf{S})^{-1}\right)$, where the structure of $\mathbf{S}$ is

$$\mathbf{S} = \begin{pmatrix} \boldsymbol{\eta}^{*T}\boldsymbol{\eta}^* & \boldsymbol{\eta}^{*T}\boldsymbol{\beta}^* & \boldsymbol{\eta}^{*T}\boldsymbol{\kappa}^* \\ \boldsymbol{\beta}^{*T}\boldsymbol{\eta}^* & \boldsymbol{\beta}^{*T}\boldsymbol{\beta}^* & \boldsymbol{\beta}^{*T}\boldsymbol{\kappa}^* \\ \boldsymbol{\kappa}^{*T}\boldsymbol{\eta}^* & \boldsymbol{\kappa}^{*T}\boldsymbol{\beta}^* & \boldsymbol{\kappa}^{*T}\boldsymbol{\kappa}^* \end{pmatrix}.$$

  The quantities of $\boldsymbol{\eta}^*$, $\boldsymbol{\beta}^*$, and $\boldsymbol{\kappa}^*$ are defined below. Let $k(i)$ denote the subpopulation for animal i. Suppose animals 1, 2, and n are from the 2nd, Kth and Jth subpopulations, respectively, then $k(1) = 2$, $k(2) = K$, and $k(n) = J$. The population parameters corresponding to all animals are $\boldsymbol{\eta}_0 = (\eta_{0k(1)}\ \eta_{0k(2)}\ \ldots\ \eta_{0k(n)})$, $\boldsymbol{\beta}_0 = (\beta_{0k(1)}\ \beta_{0k(2)}\ \ldots\ \beta_{0k(n)})$, and $\boldsymbol{\kappa}_0 = (\kappa_{0k(1)}\ \kappa_{0k(2)}\ \ldots\ \kappa_{0k(n)})$. The individual parameter vectors are $\boldsymbol{\eta} = (\eta_1\ \eta_2\ \ldots\ \eta_n)$, $\boldsymbol{\beta} = (\beta_1\ \beta_2\ \ldots\ \beta_n)$, and $\boldsymbol{\kappa} = (\kappa_1\ \kappa_2\ \ldots\ \kappa_n)$. Then $\boldsymbol{\eta}^* = \boldsymbol{\eta} - \boldsymbol{\eta}_0 = (\eta_1 - \eta_{0k(1)}\ \ \eta_2 - \eta_{0k(2)}\ \ \ldots\ \ \eta_n - \eta_{0k(n)})^T$, $\boldsymbol{\beta}^* = \boldsymbol{\beta} - \boldsymbol{\beta}_0$, and $\boldsymbol{\kappa}^* = \boldsymbol{\kappa} - \boldsymbol{\kappa}_0$ are the differences between individual parameters and the corresponding subpopulation averages.

- $p(\boldsymbol{\theta}_{0k} \mid \mathbf{y}, \boldsymbol{\theta}, \mathbf{G}, \sigma_e^2) \sim N(\bar{\boldsymbol{\theta}}_k, \mathbf{G}/n_k)$, where $\bar{\boldsymbol{\theta}}_k = \sum_{i\ s.t.\ k(i)=k} \boldsymbol{\theta}_i/n_k$ is the average of all animals associated with the k^{th} subpopulation of size n_k, $k = 1, 2, \ldots, m$.

- $p(\boldsymbol{\theta}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_{0k(i)}, \mathbf{G}, \sigma_e^2) = p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{0k(i)}, \mathbf{G}) p(\mathbf{y}_i \mid \boldsymbol{\theta}_i, \sigma_e^2)$

$$\propto\ exp\left\{\frac{-1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)})^T\mathbf{G}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)})\right\}$$
$$exp\left\{\frac{-1}{2\sigma_e^2}\sum_{j=1}^{r_i}(y_{ij} - f(\boldsymbol{\theta}_i, t_{ij}))^2\right\} \qquad (6.3)$$

Note that the posterior distribution for $\boldsymbol{\theta}_i$ does not follow any known parametric family. However, if we are willing to consider the single parameter η_i (the upper asymptote) conditional on all others, then η_i follows a normal distribution.

$p(\eta_i \mid \mathbf{y}, \beta_i, \kappa_i, \boldsymbol{\theta}_{0k(i)}, \mathbf{G}, \sigma_e^2)$

$$\propto\ exp\left\{\frac{-1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)})^T\mathbf{G}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)})\right\} exp\left\{\frac{-1}{2\sigma_e^2}\sum_{j=1}^{r_i}\{y_{ij} - f(\boldsymbol{\theta}_i, t_{ij})\}^2\right\}$$

$$\propto \quad exp\left\{\frac{-1}{2}\left((\eta_i-\eta_{0k(i)})^2G^{\eta\eta}-2(\eta_i-\eta_{0k(i)})(\beta_i-\beta_{0k(i)})G^{\eta\beta}\right.\right.$$

$$\left.\left.-2(\eta_i-\eta_{0k(i)})(\kappa_i-\kappa_{0k(i)})G^{\eta\kappa}\right)\right\}\ exp\left\{\frac{-1}{2\sigma_e^2}\sum_{j=1}^{r_i}(y_{ij}-\eta_iQ_{ij})^2\right\}.$$

$$\text{where } Q_{ij}=exp\left\{-exp\left(-\beta_i\left(\frac{t_{ij}}{\kappa_i}-1\right)\right)\right\}$$

$$\propto \quad exp\left\{\frac{-1}{2}\left((G^{\eta\eta}+\sum_{j=1}^{r_i}Q_{ij}^2/\sigma_e^2)\eta_i^2-2\left(\sum_{j=1}^{r_i}y_{ij}Q_{ij}/\sigma_e^2+\eta_{0k(i)}G^{\eta\eta}\right.\right.\right.$$

$$\left.\left.\left.-(\beta_i-\beta_{0k(i)})G^{\eta\beta}-(\kappa_i-\kappa_{0k(i)})G^{\eta\kappa}\right)\eta_i\right)\right\}$$

$$\propto \quad \sim N(\eta_i\mid\hat{\eta}_i,\hat{\mathbf{\Sigma}}_{\eta_i}).$$

where $\hat{\mathbf{\Sigma}}_{\eta_i}=(\sum_{j=1}^{r_i}Q_{ij}^2/\sigma_e^2+G^{\eta\eta})^{-1}$.

$$\hat{\eta}_i=\hat{\mathbf{\Sigma}}_{\eta_i}\left[\sum_j y_{ij}Q_{ij}/\sigma_e^2+\eta_{0k(i)}G^{\eta\eta}-(\beta_i-\beta_{0k(i)})G^{\eta\beta}-(\kappa_i-\kappa_{0k(i)})G^{\eta\kappa}\right].$$

Unfortunately, the parameters β_i and κ_i in the full conditional posterior distribution of $\boldsymbol{\theta}_i$ do not have recognizable posterior distributions. Therefore, M-H steps are required to implement Gibbs sampling to generate β_i and κ_i for this model.

6.3 Metropolis-Hastings algorithms

The Metropolis-Hastings (M-H) algorithm is commonly used when it is impossible to sample directly from a distribution of interest, e.g., the full posterior distribution of $\boldsymbol{\theta}_i$ in the previous section. To implement a M-H algorithm, a jumping distribution is required to generate a candidate value for the next state for the Markov chain. Then the magnitude of the ratio of importance ratios determines the acceptance or rejection of the candidate points. Hence, the choice of jumping distributions may change the course of a simulation and, therefore, affect the efficiency of MCMC.

For convenience, additional notation is defined here before introducing the various M-H algorithms adopted here.

Notation

$\delta_{ij}=y_{ij}-f(\boldsymbol{\theta}_i,t_{ij})$= the residual for y_{ij}.

$J(\boldsymbol{\theta}_i \mid \tilde{\boldsymbol{\theta}}_i)$: a jumping distribution given the information at point $\tilde{\boldsymbol{\theta}}_i$.

$\boldsymbol{\theta}_i^{MLE}$: the maximum likelihood estimate for $\boldsymbol{\theta}_i$ using the data only from animal i

$\boldsymbol{\theta}_i^c$: the current value for $\boldsymbol{\theta}_i$ at a certain iteration of the Markov chain.

$\boldsymbol{\theta}_i^*$: a candidate value for $\boldsymbol{\theta}_i$, typically drawn from the jumping distribution.

$\tilde{\boldsymbol{\theta}}_i$ (or $\tilde{\boldsymbol{\theta}}_i^c$, or $\tilde{\boldsymbol{\theta}}_i^*$): the mean of a normal jumping distribution derived by linearization of the Gompertz function at $\boldsymbol{\theta}_i$ (or $\boldsymbol{\theta}_i^c$, or $\boldsymbol{\theta}_i^*$).

c : a constant used to adjust the size of the variance matrix of the jumping distribution in order to allow the acceptance rate to be between 0.23 and 0.45.

$\mathbf{f}_{ij}$= the Gompertz function gradient for animal i at time $t_{ij} = \left(\frac{\partial f(\boldsymbol{\theta}_i, t_{ij})}{\partial \eta_i} \; \frac{\partial f(\boldsymbol{\theta}_i, t_{ij})}{\partial \beta_i} \; \frac{\partial f(\boldsymbol{\theta}_i, t_{ij})}{\partial \kappa_i}\right)^T =$ $(Q_{ij} \; R_{ij} \; W_{ij})^T$, where

$$
\begin{aligned}
Q_{ij} &= \frac{\partial f(\boldsymbol{\theta}_i, t_{ij})}{\partial \eta_i} = exp(-exp(-\beta_i(t_{ij}/\kappa_i - 1))) \\
R_{ij} &= \frac{\partial f(\boldsymbol{\theta}_i, t_{ij})}{\partial \beta_i} = f(\boldsymbol{\theta}_i, t_{ij}) exp(-\beta_i(t_{ij}/\kappa_i - 1))(t_{ij}/\kappa_i - 1) \\
W_{ij} &= \frac{\partial f(\boldsymbol{\theta}_i, t_{ij})}{\partial \kappa_i} = f(\boldsymbol{\theta}_i, t_{ij})(-(exp(-\beta_i(t_{ij}/\kappa_i - 1)))(\beta_i t_{ij}/\kappa_i^2)
\end{aligned}
$$

$l_i = \frac{-1}{2\sigma_\varepsilon^2} \sum_{j=1}^{r_i} (y_{ij} - f(\boldsymbol{\theta}_i, t_{ij}))^2$= the contribution to the log-likelihood from the ith animal's data.

$\mathbf{F}_i$= the gradient of the log-likelihood for animal $i = \left(\frac{\partial l_i}{\partial \eta_i} \; \frac{\partial l_i}{\partial \beta_i} \; \frac{\partial l_i}{\partial \kappa_i}\right)^T = \left(F_i^\eta \; F_i^\beta \; F_i^\kappa\right)^T$, where

$F_i^\eta = \frac{1}{\sigma_\varepsilon^2} \sum_{j=1}^{r_i} \delta_{ij} Q_{ij}$. $F_i^\beta = \frac{1}{\sigma_\varepsilon^2} \sum_{j=1}^{r_i} \delta_{ij} R_{ij}$. $F_i^\kappa = \frac{1}{\sigma_\varepsilon^2} \sum_{j=1}^{r_i} \delta_{ij} W_{ij}$.

$\mathbf{H}_i$ = the second derivative matrix of the log-likelihood for animal i

$$
= \begin{pmatrix}
\frac{\partial^2 l_i}{\partial \eta_i^2} & \frac{\partial^2 l_i}{\partial \eta_i \partial \beta_i} & \frac{\partial^2 l_i}{\partial \eta_i \partial \kappa_i} \\
\frac{\partial^2 l_i}{\partial \eta_i \partial \beta_i} & \frac{\partial^2 l_i}{\partial \beta_i^2} & \frac{\partial^2 l_i}{\partial \beta_i \partial \kappa_i} \\
\frac{\partial^2 l_i}{\partial \eta_i \partial \kappa_i} & \frac{\partial^2 l_i}{\partial \beta_i \partial \kappa_i} & \frac{\partial^2 l_i}{\partial \kappa_i^2}
\end{pmatrix}
$$

$$= \frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial Q_{ij}}{\eta_i} - Q_{ij}^2 & \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial Q_{ij}}{\partial \beta} - Q_{ij} R_{ij} & \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial Q_{ij}}{\partial \kappa} - Q_{ij} W_{ij} \\ \sum_{j=1}^{nri} \delta_{ij} \frac{\partial R_{ij}}{\partial \eta} - Q_{ij} R_{ij} & \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial R_{ij}}{\beta_i} - R_{ij}^2 & \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial R_{ij}}{\partial \kappa} - R_{ij} W_{ij} \\ \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial W_{ij}}{\partial \eta} - Q_{ij} W_{ij} & \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial W_{ij}}{\partial \beta} - R_{ij} W_{ij} & \sum_{j=1}^{r_i} \delta_{ij} \frac{\partial W_{ij}}{\kappa_i} - W_{ij}^2 \end{pmatrix}$$

$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i}$ (or $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i^*}$. or $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i^{MLE}}$)= the estimate of the variance matrix evaluated at $\boldsymbol{\theta}_i$ (or $\boldsymbol{\theta}_i^*$. or $\boldsymbol{\theta}_i^{MLE}$). The estimate can be developed from the Hessian matrix $[-\mathbf{H}_i]^{-1}$ or from the product of the gradient vector $\left[\mathbf{F}_i \mathbf{F}_i^T\right]^{-1}$.

6.3.1 Jumping distributions and linearization

Our basic computational approach is the same as for the linear mixed model. that is. using the Gibbs sampler to explore the joint posterior distribution. We have already noted that $\boldsymbol{\theta}_i$ can't be drawn directly from its full conditional posterior distribution. so that a M-H algorithm is required for that Gibbs sampling steps. In our study. we use three approaches to obtain a normal jumping distribution for the M-H algorithms to sample from the full conditional distribution of $\boldsymbol{\theta}_i$ in (6.3).

I. Random walk M-H : The first approach uses a normal distribution with mean equal to the current value and variance matrix equal to the inverse of the negative of the Hessian matrix at $\boldsymbol{\theta}_i^{MLE}$. The resulting Markov chain is a random walk about the parameter space. Note that the jumping distribution does not attempt to approximate the shape of the posterior distribution.

II. Normal approximation to the likelihood : The second approach is to construct a normal approximation to the likelihood function by linearizing the Gompertz function around the current value $\boldsymbol{\theta}_i^c$. As described in Section 4.4.1. the nonlinear function is expanded in a first-order Taylor series. The mean and variance of the jumping distribution are derived as follows:

$$
\begin{aligned}
& J(\theta_i \mid \mathbf{y}, \theta_i^c, \boldsymbol{\theta}_{-i}, \boldsymbol{\theta}_{0k(i)}, \mathbf{G}, \sigma_e^2) \\
& \propto \; exp\left\{ \frac{-1}{2\sigma_e^2} \sum_{j=1}^{r_i} \left(y_{ij} - f(\boldsymbol{\theta}_i^c, t_{ij}) - \left. \frac{\partial f}{\partial \boldsymbol{\theta}_i} \right|_{\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^c} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^c) \right)^2 \right\} \\
& \propto \; exp\left\{ \frac{-1}{2\sigma_e^2} \sum_{j=1}^{r_i} (y_{ij} - f(\boldsymbol{\theta}_i^c, t_{ij}) - Q_{ij}(\eta_i - \eta_i^c) - R_{ij}(\beta_i - \beta_i^c) - W_{ij}(\kappa_i - \kappa_i^c))^2 \right\} \\
& \sim \; N(\bar{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i}). \qquad (6.4)
\end{aligned}
$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i} = (\sum_j^{r_i} \mathbf{f}_{ij} \mathbf{f}_{ij}^T)^{-1} \sigma_e^2$. $\quad \bar{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i^c + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i} \mathbf{F}_i$

Alternative versions arise if we use the Hessian matrix to produce a second derivative estimator, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i} = -\mathbf{H}_i^{-1} \sigma_e^2$, or by altering the point at which the linearization is done.

III. Normal approximation to the posterior distribution: The third approach is similar to the second approach, except that a normal approximation is developed for the posterior distribution rather than just for the likelihood function:

$$
\begin{aligned}
& J(\boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{\theta}_i^c, \boldsymbol{\theta}_{-i}, \boldsymbol{\theta}_{0k(i)}, \mathbf{G}, \sigma_e^2) \\
& \propto \; exp\left\{ \frac{-1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)})^T \mathbf{G}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)}) \right\} \\
& \qquad exp\left\{ \frac{-1}{2\sigma_e^2} \sum_{j=1}^{r_i} \left(y_{ij} - f(\boldsymbol{\theta}_i^c, t_{ij}) - \left. \frac{\partial f}{\partial \boldsymbol{\theta}_i} \right|_{\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^c} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^c) \right)^2 \right\} \\
& \propto \; exp\left\{ \frac{-1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)})^T \mathbf{G}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0k(i)}) \right\} \\
& \qquad exp\left\{ \frac{-1}{2\sigma_e^2} \sum_{j=1}^{r_i} (y_{ij} - f(\boldsymbol{\theta}_i^c, t_{ij}) - Q_{ij}(\eta_i - \eta_i^c) - R_{ij}(\beta_i - \beta_i^c) - W_{ij}(\kappa_i - \kappa_i^c))^2 \right\} \\
& \sim \; N(\bar{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i}). \qquad (6.5)
\end{aligned}
$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i} = (\sum_j^{r_i} \mathbf{f}_{ij} \mathbf{f}_{ij}^T / \sigma_e^2 + \mathbf{G}^{-1})^{-1}$. $\quad \bar{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i^c + (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i}) \left\{ \mathbf{F}_i + \mathbf{G}^{-1} \boldsymbol{\theta}_{0k(i)} \right\}$.

When the prior distribution of $\boldsymbol{\theta}_i$ has a large variance ($\mathbf{G}$ is large) then approaches II and III are similar. Also, when the data dominate the posterior distribution (e.g., if the sample size is large), then the two will be similar. For approach II, the prior distribution

does not contribute to the derivation of the jumping distribution for the selection of candidates, but the prior distribution does play a role in the ratio of importance ratios for the Metropolis-Hastings algorithm.

$$\alpha = \frac{\mathbf{p}(\boldsymbol{\theta}_i^* \mid \mathbf{y})/J(\boldsymbol{\theta}_i^* \mid \boldsymbol{\theta}_i^c)}{\mathbf{p}(\boldsymbol{\theta}_i^c \mid \mathbf{y})/J(\boldsymbol{\theta}_i^c \mid \boldsymbol{\theta}_i^*)} = \frac{\mathbf{p}(\mathbf{y} \mid \boldsymbol{\theta}_i^*)\mathbf{p}(\boldsymbol{\theta}_i^*)/N(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\theta}}_i^c, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i^c})}{\mathbf{p}(\mathbf{y} \mid \boldsymbol{\theta}_i^c)\mathbf{p}(\boldsymbol{\theta}_i^c)/N(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\theta}}_i^*, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_i^*})}$$

For approach III, the prior distribution plays a role in selecting candidate values, but it is canceled out in α and consequently has no contribution to the size of α.

6.3.2 Specific Metropolis-Hastings algorithms

The previous subsection described approaches for developing jumping distributions. Here we describe the specific algorithms that are applied in our study. M-H algorithms for sampling $\boldsymbol{\theta}_i$ are discussed by Bennett et al. (1996) and Gilk and Roberts (1996).

Algorithm 1 (A1): Random walk M-H algorithm

$J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c) = N(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_i^{MLE}})$. The mean of the jumping distribution is set at the current $\boldsymbol{\theta}_i^c$ and its variance is determined by the curvature around $\boldsymbol{\theta}_i^{MLE}$. Note that the variance of the jumping distribution for each $\boldsymbol{\theta}_i$ only needs to be calculated once. Also the jumping distribution is symmetric, so that $J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c) = J(\boldsymbol{\theta}_i^c \mid \boldsymbol{\theta}_i)$, and therefore the format of the ratio used to determine acceptance of the candidates acceptance rate can be simplified.

Algorithm 2 (A2): Scale-dependent random walk M-H algorithm

$J(\boldsymbol{\theta}_i \mid \tilde{\boldsymbol{\theta}}_i) = N(\boldsymbol{\theta}_i \mid \tilde{\boldsymbol{\theta}}_i, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_i})$. The jumping distribution is derived by linearization of the likelihood alone, its mean is set at $\tilde{\boldsymbol{\theta}}_i$ (either $\boldsymbol{\theta}_i^c$ or $\boldsymbol{\theta}_i^*$), and its variance is determined by the curvature around $\tilde{\boldsymbol{\theta}}_i$. Note that it is not necessary for the jumping distribution to be symmetric, since the variance depends on $\tilde{\boldsymbol{\theta}}_i$.

Algorithm 3 (A3): Mean-scale-dependent M-H algorithm

$J(\boldsymbol{\theta}_i \mid \tilde{\boldsymbol{\theta}}_i) = N(\boldsymbol{\theta}_i \mid \bar{\tilde{\boldsymbol{\theta}}}_i, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_i})$. The jumping distribution is derived by linearization

of the likelihood alone, and its mean and variance are determined by linearization around $\tilde{\boldsymbol{\theta}}_i$ (see (6.4)). Note that the only difference between A3 and A2 is the location of the mean for the jumping distribution.

Algorithm 4 (A4): Local-mode-dependent M-H algorithm

$J(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^c) = N(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_i^{Mode}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_i^{Mode}})$. The jumping distribution is derived from the joint posterior distribution rather than from the likelihood alone. During the random walk along the target surface, the jumping distribution is evaluated at the local mode ($\boldsymbol{\theta}_i^{Mode}$), which is estimated either by a Newton-Raphson method or a Gauss-Newton method with the current point $\boldsymbol{\theta}_i^c$ as the starting point. Note that the mean and variance of the jumping distribution are determined by $\boldsymbol{\theta}_i^{Mode}$.

6.3.3 Batching

Another way in which M-H algorithms can vary is in whether the elements of $\boldsymbol{\theta}_i$ are treated individually or batched together in a single M-H algorithm (as assumed in Section 6.3.2). This is relevant in the Gompertz model because, as noted earlier, the full conditional distribution for η_i is a normal distribution. We consider three batching schemes.

Scheme 1 (B1): draw η_i, β_i, and κ_i one-by-one in sequence.

Scheme 2 (B2): draw η_i from its full conditional posterior distribution and then draw vector $(\beta_i\ \kappa_i)$, and

Scheme 3 (B3): draw the vector $(\eta_i\ \beta_i\ \kappa_i)$ in a single trivariate M-H step.

Algorithms incorporating Scheme B3 are those described in Section 6.3.2. For Scheme B2, the full conditional posterior distribution of $(\beta_i\ \kappa_i)$ involves the nonlinear Gompertz function, so linearization is required to derive a jumping distribution for sampling candidate points. This approach is the same as in the previous section (see (6.4) and (6.5))

except that only β_i and κ_i are incorporated in the distribution. For Scheme B1, the full conditional posterior distributions of β_i and κ_i are required. These take similar forms, so only the one for β_i is shown here:

$$J(\beta_i \mid \mathbf{y}, \beta_i^c, \eta_i^c, \kappa_i^c, \sigma_e^2) \sim N(\hat{\beta}_i, \hat{\mathbf{\Sigma}}_{\beta_i}),$$

with

$$\hat{\beta}_i = \beta_i^c + (\sum_{j=1}^{r_i} R_{ij}^2)^{-1} \sum_{j}^{r_i} R_{ij}(y_{ij} - f(y_{ij} \mid \boldsymbol{\theta}_i^c, \sigma_e^2)), \text{ and}$$

$$\hat{\mathbf{\Sigma}}_{\beta_i} = (\sum_{j=1}^{r_i} R_{ij}^2)^{-1}\sigma_e^2, \text{ where } R_{ij} = \left.\frac{\partial f(\boldsymbol{\theta}_i)}{\partial \beta_i}\right|_{\beta_i=\beta_i^c}$$

Besides applying different batching schemes to $\boldsymbol{\theta}_i$ within the M-H algorithms, batching schemes can also be applied to the hyperparameters $\boldsymbol{\theta}_{0k}$ $(k = 1, \ldots, m)$, whose posterior distribution follows a multivariate normal distribution (see Section 6.2.2). When drawing η_{0k}, β_{0k}, and κ_{0k} in sequence, the full conditional posterior distribution of β_{0k} becomes $N(\hat{\beta}_0, \hat{\mathbf{\Sigma}}_{\beta_0})$, with

$$\hat{\beta}_0 = \sum_{i \text{ s.t. } k(i)=k}^{n_k} \beta_i/n_k + (n_k G^{\beta\beta})^{-1}\left[G^{\eta\beta}\sum_{i \text{ s.t. } k(i)=k}^{n_k}(\eta_i - \eta_{0k(i)}) + G^{\beta\kappa}\sum_{i \text{ s.t. } k(i)=k}^{n_k}(\kappa_i - \kappa_{0k(i)})\right],$$

$$\text{and } \hat{\mathbf{\Sigma}}_{\beta_0} = (n_k G^{\beta\beta})^{-1},$$

where n_k is the total number of animals in subpopulation k. Similarly, the full conditional posterior distribution of η_{0k} and κ_{0k} can be derived by substituting the relevant elements in the above expression. One would expect Scheme B3 to be best for the hyperparameters, since batching is generally effective and the full conditional posterior distribution is a known Gaussian (rather than requiring an M-H approximation).

The M-H algorithms and batching schemes used in this chapter are summarized in Table 6.1. The means and variances of the normalized jumping distributions and the ratios of importance ratios for these M-H algorithms are listed in Table 6.2. Note that only batching scheme B3 is used with algorithms A1 and A4.

Table 6.1 Notation and description for the Metropolis-Hastings algorithms for the nonlinear model.

Algorithm	Description
A1	**Random walk M-H algorithm:** The mean of the jumping distribution is located at the current value $\boldsymbol{\theta}_i^c$ and its variance is determined by the curvature around the MLE of $\boldsymbol{\theta}_i$
A2	**Scale-dependent random walk M-H algorithm:** The mean of the jumping distribution is located at $\tilde{\boldsymbol{\theta}}_i^{(1)}$ and its variance is determined by the variance of the linear approximate likelihood at $\tilde{\boldsymbol{\theta}}_i$
A3	**mean-scale-dependent M-H algorithm:** The mean and variance of the jumping distribution are determined by the point where it is evaluated, say $\tilde{\boldsymbol{\theta}}_i$
A4	**local-mode-dependent M-H algorithm:** The mean and variance of the jumping distribution are determined by the local mode, say $\boldsymbol{\theta}_i^{Mode}$, which is updated by the Newton-Raphson or Gauss-Newton method at each iteration, given $\boldsymbol{\theta}_i^c$
B1	draw $\eta_i, \beta_i, \kappa_i$ one-by-one in sequence
B2	draw η_i and then (β_i, κ_i) in vector form
B3	draw $(\eta_i, \beta_i, \kappa_i)$ in vector form

(1):$\tilde{\boldsymbol{\theta}}_i$ can be the current point $\boldsymbol{\theta}_i^c$ or the candidate point $\boldsymbol{\theta}_i^*$

6.4 Comparison of Metropolis-Hastings algorithms

The efficiency of an algorithm is expressed in terms of the number of iterations required until convergence is diagnosed, denoted by γ. The smaller the number of iterations required, the more efficient is the algorithm. For convergence diagnosis, several parallel Markov chains, with overdispersed starting points, are simulated, and the second halves of the chains are used to calculate convergence statistics. The multivariate potential scale reduction (MPSR, Gelman and Rubin, 1998), based on comparing between-sequence and within-sequence variances, is used here as the convergence diagnostic. If the estimate of MPSR, $\sqrt{R^p}$, is below 1.2, then we consider that the chains have converged. In practice, we determine the convergence point γ by computing the diagnostic every s iterations.

Table 6.2 The mean and variance of the normal jumping distribution and the ratio of importance ratios α for each Metropolis-Hastings algorithm ($\tilde{\boldsymbol{\theta}}_t$ can be the current point $\boldsymbol{\theta}_t^c$ or the candidate point $\boldsymbol{\theta}_t^*$).

M-H Algorithm[1]	Batch[1]	Jumping Distn. Mean	Jumping Distn. Variance	$\alpha = \frac{p(\boldsymbol{\theta}_t^*\|\mathbf{y})/J(\boldsymbol{\theta}_t^*\|\boldsymbol{\theta}_t^c)}{p(\boldsymbol{\theta}_t^c\|\mathbf{y})/J(\boldsymbol{\theta}_t^c\|\boldsymbol{\theta}_t^*)}$
A1	B3	$\boldsymbol{\theta}_t^c$	$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{MLE}$	$\frac{p(\boldsymbol{\theta}_t^*\|\mathbf{y})}{p(\boldsymbol{\theta}_t^c\|\mathbf{y})}$
A2	B1-B3	$\tilde{\boldsymbol{\theta}}_t$	$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_t}$	$\frac{p(\boldsymbol{\theta}_t^*\|\mathbf{y})/N(\boldsymbol{\theta}_t^*\|\boldsymbol{\theta}_t^c,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^c})}{p(\boldsymbol{\theta}_t^c\|\mathbf{y})/N(\boldsymbol{\theta}_t^c\|\boldsymbol{\theta}_t^*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*})}$
A3	B1-B3	$\bar{\tilde{\boldsymbol{\theta}}}_t$	$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_t}$	$\frac{p(\boldsymbol{\theta}_t^*\|\mathbf{y})/N(\boldsymbol{\theta}_t^*\|\bar{\boldsymbol{\theta}}_t^c,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^c})}{p(\boldsymbol{\theta}_t^c\|\mathbf{y})/N(\boldsymbol{\theta}_t^c\|\boldsymbol{\theta}_t^*,\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^*})}$
A4	B3	$\boldsymbol{\theta}_t^{mode}$	$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^{mode}}$	$\frac{p(\boldsymbol{\theta}_t^*\|\mathbf{y})/N(\boldsymbol{\theta}_t^*\|\boldsymbol{\theta}_t^{mode},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^{mode}})}{p(\boldsymbol{\theta}_t^c\|\mathbf{y})/N(\boldsymbol{\theta}_t^c\|\boldsymbol{\theta}_t^{mode},\boldsymbol{\Sigma}_{\boldsymbol{\theta}_t^{mode}})}$

(1): Refer to Table 6.1 for description.

For example, if s=200, the sequential values of $\sqrt{R^p}$ are calculated from the second halves of the first 200 simulated values from each chain and then the second half of the first 400 simulated values from each chain, ... etc. The convergence point is said to be γ (a multiple of s), when $\sqrt{R^p}$ is less than 1.2 for the diagnosis carried out at iteration γ and remains below 1.2 for 20,000 iterations after that. This avoids cases where the MCMC algorithm appears to have converged but has not.

6.4.1 Simulation study

In Section 6.5, we apply our various algorithms to the pig weight gain data that were introduced in Section 5.5. One difficulty then is that the true model is unknown. To better understand the performance of our algorithms, we begin by analyzing data simulated under the Gompertz model (see Sections 6.1 and 6.2). Data are generated for 1000 independent animals, 500 of each gender, with measuring times every 5 days from day 0 to day 110. There are two subpopulations (i.e., females and males). Note that genetic relationships among animals are ignored here. The parameters used to simulate data are as follows:

- $\epsilon_{ij} \sim N(0, \sigma_e^2 = 4.0)$
- Population parameters: $\boldsymbol{\theta}_{0_M} = (98.0 \;\; 1.2 \;\; 41.0)^T$ for males,
  $\boldsymbol{\theta}_{0_F} = (105.0 \;\; 1.26 \;\; 38.0)^T$ for females
- $\boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_{0k(i)}, \mathbf{G})$, where $\mathbf{G} = \begin{pmatrix} 150.0 & 0.0 & 58.0 \\ 0.0 & 0.003 & 0.0 \\ 58.0 & 0.0 & 35.0 \end{pmatrix}$.

Four M-H algorithms (A1-A4) incorporated with three batching schemes (B1-B3) are used to draw posterior samples for $\boldsymbol{\theta}_i$ (i=1, ... 1000). Three parallel Markov chains are simulated. Because some algorithms are sensitive to variation in the starting points, several starting points are tried for all algorithms.

6.4.2 Simulation results

First, we check to see whether algorithms converge and if so, whether they converge to values approximately equal to the true values that are used for simulating the data set. Table 6.3 lists the convergence point of all algorithms for a single simulated data set. Note that Algorithm A3 did not converge.

Table 6.3 Convergence point[1] for Metropolis-Hastings algorithms used for fitting nonlinear Gompertz models to a simulated data set.

Variance Type[2]	Batch[3]	M-H Algorithm[3]			
		A1	A2	A3	A4
	B1		7.200	> 80.000	
F	B2		5.600	> 80.000	
F	B3		1.600	> 80.000	800
H	B2		7.200	> 80.000	
H	B3	1.600	2.800	> 80.000	

(1) indicated by the value of γ . defined in Table 5.2
(2) F and H stand for algorithms that use the gradient- or Hessian-type matrix as approximate variance matrix
(3) A1-A4 and B1-B3 refer to methods defined in Table 6.1

Among the feasible algorithms. Table 6.4 presents the posterior means and standard deviations of population parameters and variance parameters ($\mathbf{G}$ and σ_e^2) obtained by algorithms A1B3. A2B2 and A2B1. Comparing with the column labeled true value. Table 6.4 shows that these three algorithms provide similar answers. except that for Scheme B1 the parameters related to growth rate β (e.g.. β_M. $G_{\eta\beta}$) have a larger means and larger variances. even though the diagnostic indicates that Algorithm A2B1 has converged. Table 6.4 indicates that a data set of 1000 animals is sufficiently large to identify model parameters. since at convergence all population and variance parameters approximately reflect the true values. In the reminder of this section. we compare algorithms based on convergence rate.

6.4.2.1 M-H algorithm

First we observe that Algorithm A3. in which the jumping distribution is an approximation to the likelihood function based on a linearization of the nonlinear Gompertz function at the current parameter value. does not converge within a tolerable number of

Table 6.4 Posterior means and standard deviations of some selected parameters for three M-H algorithms when fitting the nonlinear mixed model to a simulated data set.

Parameter	True value	A2B1 mean	A2B1 sd	A2B2 mean	A2B2 sd	A1B3 mean	A1B3 sd
η_M	98.0	97.70	0.57	97.72	0.57	97.73	0.57
β_M	1.2	1.21	0.004	1.20	0.003	1.20	0.003
κ_M	41.0	41.22	0.27	41.20	0.27	41.20	0.27
η_F	105.0	105.12	0.57	105.07	0.57	105.08	0.56
β_F	1.26	1.27	0.004	1.26	0.003	1.26	0.003
κ_F	38.0	38.41	0.27	38.37	0.27	38.37	0.27
G_η	150.0	154.01	7.38	152.72	7.31	152.62	7.30
G_β	0.003	0.0078	.0004	0.0028	0.0002	0.0028	0.0002
G_κ	35.0	35.14	1.66	35.14	1.67	35.1334	1.6643
$G_{\eta,\beta}$	0	-0.116	0.044	0.007	0.029	0.006	0.028
$G_{\eta,\kappa}$	58.0	59.72	3.20	59.44	3.20	59.42	3.19
$G_{\beta,\kappa}$	0	-0.052	0.021	-0.016	0.014	-0.016	0.014
σ_e^2	4.0	4.09	0.04	4.09	0.04	4.09	0.04
$\sqrt{R^p}^{(1)}$		7.200		1.600		1.600	

(1) convergence point γ, defined in Table 5.2.

iterations (i.e., convergence is not reached before 80,000 iterations, Table 6.3). Algorithm A2, in which the jumping distribution is centered at the current point (around which the linearization is applied) and with variance based on information from the linearized approximation, does converge reliably, because it can travel through the full conditional posterior distribution of $\boldsymbol{\theta}_i$ no matter what batching scheme it is incorporated with (γ = 1.600 to 7.200). Algorithms using the gradient to form a variance estimate for the jumping distribution appear to work better than using the Hessian matrix (γ=5.600 vs 7.200 for Scheme B2, and γ=1.600 vs 2.800 for scheme B3). These algorithms are also less sensitive to variation in starting points.

Algorithm A1 (implemented with batching scheme B3 only), using the curvature

around $\boldsymbol{\theta}_i^{MLE}$ ($\boldsymbol{\Sigma}_{\boldsymbol{\theta}_i^{MLE}}$) as the variance of the jumping distribution, yields a better convergence rate than algorithm A2B3 (γ=1.600 vs 2.800). In addition, A1 is easier to program and better tolerates variation in starting points. Algorithm A4, which determines an approximation to the jumping distribution by finding the local mode of the posterior distribution, yields the fastest convergence rate among these four M-H algorithms, but starting points for it need to be set in a narrow range.

6.4.2.2 Batching Scheme

Two batching schemes for the hyperparameters $\boldsymbol{\theta}_{0k}$ (k= Female or Male) were considered: batching all three components together and a single-element scheme. There was no real difference in the convergence point for these two schemes. Therefore, all comparisons presented are made for algorithms which draw the entire vector $\boldsymbol{\theta}_{0k}$.

With respect to batching schemes B1 – B3 for the individual parameter vector $\boldsymbol{\theta}_{0k(i)}$, Algorithm A2 is the only algorithm in which we can compare the convergence rates among B1 – B3. It is therefore possible to compare the different batching schemes for this algorithm. The results (Table 6.3) show that drawing the vector ($\eta_i, \beta_i, \kappa_i$) (B3, γ=1.600 or 2.800) is preferred over drawing η_i and then vector (β_i, κ_i) (B2, γ=5.600 or 7.200), and also over drawing $\eta_i, \beta_i, \kappa_i$ in sequence (B1, γ=7.200). This result is not unexpected because η (the asymptote) and κ (the inflection point) are highly correlated (r= 0.8). In such cases, batching or reparameterization are helpful tools (Gilks and Roberts, 1996).

6.4.2.3 Starting points

To diagnose convergence, we run parallel Markov chains with different overdispersed starting points. Among all parameters, only the variance matrix **G** requires starting values, because starting values for $\boldsymbol{\theta}_i$ can be generated on their basis. From covariance matrix **G** (Table 6.4), the standard deviation (sd) is 12.2 for η_i, 5.9 for κ_i, and 0.055

for β_i over individual. The correlation between η_i and κ_i is high (r= 0.8) and therefore needs to be taken into account to avoid generating unrealistic starting values for $\boldsymbol{\theta}_i$. Hence, we vary the sd values for η_i, β_i, and κ_i, and then construct the starting points for $\mathbf{G}$ (i.e., $\mathbf{G}(1,1)$ = selected value for $sd(\eta_i)^2$)), and the starting values for η_i, β_i, and κ_i can then be generated from a trivariate normal distribution.

Compared to other algorithms, the variation of starting points for Algorithm A1 can be larger (e.g., sd=30 for η, 12 for κ, and 0.14 for β) to allow the MCMC methods work. When the Hessian matrix is used for updating candidates, it requires several trials to select variation in a narrow range (e.g., sd=9 for η). For Algorithm A2 with the gradient type of variance, the variation of starting points can be set in a moderate range (e.g., sd=21 for η). But for Algorithm A4, it requires a quite narrow range for starting points (e.g., sd=6 for η). For all possible combinations of Algorithms A1 – A4, batching schemes B1 – B3, and variance type Hessian or gradient, convergence rate is lower when the variation of starting points is large, provided of course that the algorithm converges.

6.4.3 Summary and discussion

When a nonlinear function is involved in the likelihood function, the Metropolis-Hasting (M-H) algorithm is used with the MCMC algorithm to generate posterior samples for the parameters of the nonlinear function. A jumping distribution is required to implement the M-H algorithm, and the choice of the jumping distribution affects the efficiency of the M-H algorithm. In our study, four M-H algorithms and three batching schemes are considered.

Our results suggest that convergence is fastest when the jumping distribution is centered at the current value of the individual parameter vector and its variance equals to the variance of the linear approximate likelihood either at the current value or at the individual MLE. These are Algorithms A1 and A2 in this chapter. Moreover, it is better to use a batching scheme for the correlated parameters (scheme B3). We also find

that using the gradient (i.e. by linearization) to evaluate the variance matrix for each θ_i works better than using the Hessian matrix.

Comparing algorithms A4 and A1, the faster convergence for A4 comes at a large computation price, since Algorithm A4 requires that one calculate the local MLE at each iteration. In addition, A4 is very sensitive to variation in starting points. These two disadvantages of A4 limit its use for our study.

The allowable variation in starting points for the MCMC algorithm is smaller for nonlinear models than that for linear polynomial models. This may be explained by the fact that variation for most biological characters of a species is not too large. For Algorithms A2 and A3, a large variation in starting points often results in the interruption of the simulation due to undefined $\mathbf{F}$ or $\mathbf{H}$ matrices, unrealistic candidates, or slow mixing.

Besides convergence rate of the M-H algorithms, efficiency considerations for MCMC include the ease of programming and the computation time per iteration. These factors are ignored in our study.

6.5 Application to pig weight gains

6.5.1 comparison of MCMC algorithms

The data set of pig weight gains used in this section is described in Section 5.5. Since Algorithm A3 does not efficiently yield posterior samples, only Algorithms A1, A2, A4 and possible batching schemes are used for fitting the pig weight data. Three parallel Markov chains are simulated for each algorithm. The genetic relationships among animals are ignored in this section.

The convergence points for these algorithms are shown in Table 6.5. The results are in general similar to those obtained with the simulated data. Scheme B2 does not work well here. The best algorithms are A2B3 with gradient type variance matrix and A1B3. Algorithm A4B3 is still sensitive to starting points.

6.5.2 Results of model fitting

The population parameters and variance components obtained from four M-H algorithms are listed in Table 6.6 and histograms for population parameters are shown in Figure 6.2. Results show that males grow ($\beta_M = 1.19$ $se = 0.006$) slower than females ($\beta_F = 1.21$ $se = 0.006$) and with a lower asymptotic value (η_M=99.8 $se = 1.2$ vs η_F=101.1 se =1.25), and a later inflection point (κ_M=41.3 $se = 0.56$ vs κ_F=39.8 se = 0.56) (see Figure 6.3).

Table 6.5 Convergence rate[1] for Metropolis-Hastings algorithms used for fitting nonlinear Gompertz model to pig weight gains

Variance Type[2]	Batch	M-H Algorithm[3]		
		A1	A2	A4
	B1		14.400	
F	B2		20.800	
F	B3		1.200	400
H	B2		> 80.000	
H	B3	2.400	3.600	

(1) indicated by the value of convergence point γ ($\sqrt{R^p} < 1.2$), defined in Table 5.2
(2) F and H stand for algorithms using gradient- or Hessian-type matrices as approximate variance matrix
(3) A1-A4 and B1-B3 refer to Table 6.1

It is also of interest to compare the nonlinear and linear models for fitting pig weight gains. The sum of squared residuals of posterior predicted values for the Gompertz nonlinear mixed model and for random polynomial regression model are 7511.7 and 4161.1, respectively. The reason for the fact that the linear model fits this pig weight gains data set better may be due to the experimental time period, where pigs grow linearly and have not yet reached the slowing down in growth curve, (see the sample data in Figure 5.2). The ranks of the top 10 animals for each gender obtained from the nonlinear Gom-

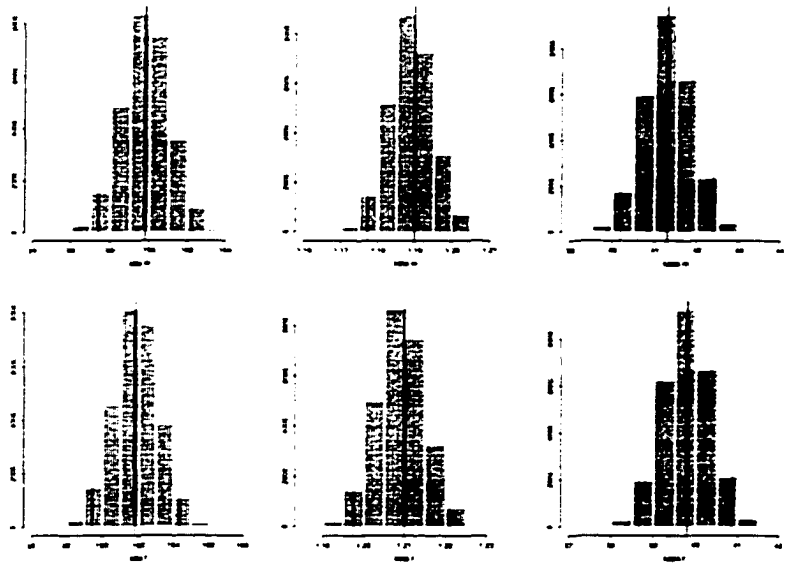

Figure 6.2 Histograms for population parameters (η, β and κ) for males and females (solid vertical line for mean, dashed vertical lines for quantiles 2.5 and 97.5).

pertz model consequently differ from those obtained from the linear random regression model M42 (ignoring relationships between animals) (see Table 5.11). Given that the linear models are easier to fit and fit better, we recommend that longitudinal data be examined initially with polynomial random regression models to determine if nonlinear models are required.

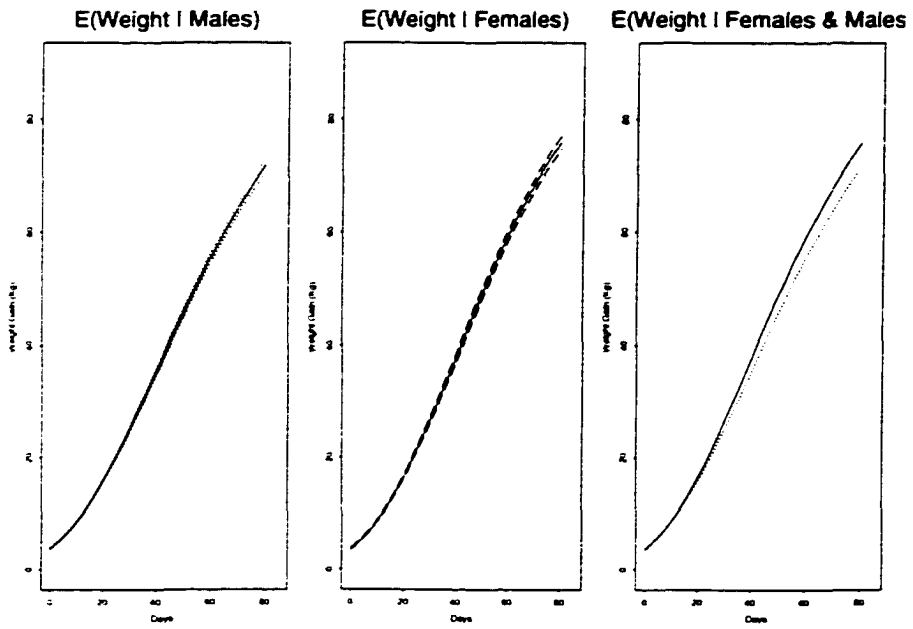

Figure 6.3 The 95% posterior region for the weight gain of each gender. left: for males (mean curve: solid line); middle: for females (mean curve: solid line): right: mean curve for females (solid line), for males (dashed line).

Table 6.6 Posterior means and standard deviations of selected parameters for four M-H algorithms used to fit pig weight gains

	A2B1		A2B3	
Parameter	**mean**	**sd**	**mean**	**sd**
η_M	99.65	1.23	99.81	1.223
β_M	1.19	0.007	1.19	0.006
κ_M	41.24	0.56	41.29	0.56
η_F	101.73	1.26	101.90	1.25
β_F	1.21	0.007	1.21	0.006
κ_F	39.73	0.56	39.77	0.562
G_η	123.09	15.26	122.51	14.861
G_β	0.004	0.0005	0.0025	0.0003
G_κ	26.28	3.04	26.125	3.05
$G_{\eta,\beta}$	0.011	0.066	0.082	0.051
$G_{\eta,\kappa}$	42.85	6.06	42.579	5.998
$G_{\beta,\kappa}$	0.009	0.029	0.030	0.024
σ_e^2	1.97	0.045	1.97	0.046
γ	14.400		3.600	
	A2B2		**A1B3**	
Parameter	**mean**	**sd**	**mean**	**sd**
η_M	99.89	1.24	99.82	1.22
β_M	1.19	0.006	1.19	0.006
κ_M	41.32	0.57	41.29	0.56
η_F	101.76	1.21	101.88	1.25
β_F	1.21	0.006	1.21	0.006
κ_F	39.72	0.55	39.76	0.56
G_η	122.06	14.55	122.35	15.00
G_β	0.0025	0.0003	0.0025	0.0003
G_κ	26.00	3.03	26.09	3.06
$G_{\eta,\beta}$	0.081	0.051	0.081	0.051
$G_{\eta,\kappa}$	42.34	5.88	42.52	6.05
$G_{\beta,\kappa}$	0.030	0.023	0.029	0.023
σ_e^2	1.97	0.045	1.97	0.046
γ	10.400		2.400	

CHAPTER 7 SUMMARY

A linear mixed model for repeated measurements such as weight gain of an animal over time is $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}$ with fixed effects $\mathbf{b}$ and animal random effects $\mathbf{u}$ besides random residuals $\boldsymbol{\epsilon}$. An alternative to the REML-BLUP approach for drawing inference is the Bayesian approach, which combines information provided by the data and prior knowledge about the model parameters to draw inferences. Generating samples from the posterior distribution often relies on Markov chain Monte Carlo (MCMC) methods. Our studies primarily focus on the efficiency of MCMC methods in models for repeated measurements in time. Efficient methods should make Bayesian methods feasible in studying animal growth traits.

One common type of model for analyzing repeated data over time are polynomial models. Such models have often been analyzed using the traditional REML-BLUP approach. When a polynomial model is able to describe the population average curve, but it is desirable to allow the coefficients of some or all polynomial terms to be affected by individual differences, then a random polynomial regression model can be applied.

With conjugate prior distributions, posterior samples under the random regression model can be obtained with Gibbs sampling algorithm. Orthogonality of parameters can reduce posterior correlations among model parameters and therefore improve convergence rate of MCMC methods. In the random regression model, this is achieved by the use of Legendre polynomials.

Hierarchical centering can improve the convergence rate of MCMC methods when the variance of the random effects is much larger than the variance of the residuals. Adopting

hierarchical centering and orthogonality simultaneously yields the greatest improvement in convergence rate. We also find that batching parameters (i.e., drawing a vector of subpopulation-level or individual-level parameters together rather than element-by-element) improve convergence.

When the additive genetic relationship among animals is incorporated, there are animal genetic and permanent environmental random components besides random residuals. Then the model is $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \epsilon$. Using a cycling algorithm along with Legendre polynomials leads to the best performance. Note that this approach does not require a choice between the two possible hierarchical centering algorithms: centering genetic effects or centering permanent environmental effects.

When fitting nonlinear random regression models, more sophisticated Metropolis-Hastings algorithms are required. Jumping distributions are used to generate candidate posterior points in these algorithms. Several algorithms are proposed based on varying the center of the jumping distribution or the variance of the jumping distribution. Linearization of nonlinear functions is a key element of our approach to deriving jumping distributions. Our results suggest that convergence is fastest when the jumping distribution is centered at the current value of the Markov chain and its variance matrix is evaluated either at the current value or at the individual-model-fitting MLE for the corresponding parameters. We also found that using the gradient to evaluate the variance matrix for each individual parameter vector is more reliable than using the Hessian matrix. Once again, it is better to batch correlated parameters than considering them one by one. In general, the range of starting points that yield reasonable convergence rate for MCMC algorithm is smaller for nonlinear models than for linear models.

In summary, the results provided here suggest that it possible to develop methods for efficient implementation of Bayesian methods in random regression models for repeated measures data as is collected to study growth of animals. In polynomial models, the use of orthogonal polynomials and hierarchical centering leads to MCMC algorithms

that converge quickly. It is also possible to develop algorithms for random regression models that are not linear in the parameters, known as nonlinear models in this thesis. We find that Metropolis-Hastings algorithms with a normal jumping distribution, that is centered at the current value and with its variance evaluated either at individual-model-fitting MLE or at the current value, perform best. Additional work is required to determine if the same type of the jumping distributions is best for different nonlinear models.

BIBLIOGRAPHY

Andersen, S., and Petersen, B. (1996). Growth and food intake curves for group-housed gilts and castrated male pigs. *Animal Science*, **63**: 457-464.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley & Sons.

Bennett, J. E., Racin-Poon A., and Wakefield, J. C. (1996). MCMC for nonlinear hierarchical models. In *Markov Chain Monte Carlo in Practice* (Eds Gilks, Richardson, and Spiegelhalter), 339–357. London: Chapman & Hall.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer-Verlag.

Besag, J., and Green P. (1993). Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B*, **55**: 25-37.

Besag, J., Green P., Higdon D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**: 3-41.

Blasco, A., Sorensen, D., and Bidandel, J. P. (1998). Bayesian inference of genetic parameters and selection response for litter size components in pigs. *Genetics*, **49**: 301-306.

Brooks, S., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Computational and Graphical Statistics*, **7**: 434-455.

Casella, G., and George, E. I. (1992). Explaining the Gibbs sampler. *J. American Stat. Assoc.*, **46**: 167-174.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**: 327-335.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics.* Essex: Longman, U. K.

Gelfand, A. and Carlin, B. (1995). Comment on: Bayesian computation and stochastic systems. *Statistical Science*, **10**: 43-46.

Gelfand, A., Sahu, S. K., and Carlin, B. P. (1995a). Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**: 479-488.

Gelfand, A., Sahu, S. K., and Carlin, B. P. (1995b). Efficient parametrisations for generalized linear mixed models. In *Bayesian Statistics 5*. Eds. Bernardo, Berger, David, and Smith, 165-180.

Gelfand, A., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. American Stat. Assoc.*, **85**: 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995a). *Bayesian Data Analysis.* London: Chapman & Hall.

Geman, A., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721-741.

Gelman, A., Roberts, G., and Gilks, W. (1995b). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*. Eds. Bernardo, Berger, David, and Smith, 599-607. London: Oxford University Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**: 457-511.

Gianola, D., and Fernando, L. R. (1986). Bayesian methods in animal breeding theory. *J. Animal Science*, **63**: 217-244.

Gilks, W. R. (1996). Full conditional distributions. In *Markov Chain Monte Carlo in Practice* Eds. Gilks, Richardson, and Spiegelhalter, 75-88. London: Chapman & Hall.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996). Introduction to Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*. Eds. Gilks, Richardson, and Spiegelhalter, 1-20. London: Chapman & Hall.

Gilks, W. R., and Roberts, G. O. (1996). Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice*. Eds. Gilks, Richardson, and Spiegelhalter, 89-114. London: Chapman & Hall.

Gilmour, A. R., Thompson, R., and Cullis B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**: 1440-1450.

Gildberger, A. S. (1962). Best linear unbiased prediction in the generalised linear regression model. *J. American Stat. Assoc.*, **57**: 369-375.

Graser, H. U., Smith, S. P., and Tier, B. (1987). A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *J. Animal Science*, **64**: 1362-1370.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**: 97-109.

Harville, D. A. (1975). Maximum likelihood approaches to variance component estimation and to related problems. Technical Report No. 75-0175, Aerospace Research Laboratories, Wright-Patterson, AFB, Ohio.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics*, **4**: 384-395.

Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and to related problems. *J. American Stat. Assoc.*, **72**: 320-338.

Harville, D. A. (1985). Decomposition of prediction error. *J. American Stat. Assoc.*, **80**: 132-138.

Harville, D. A. (1991) Comment on Robinson, That BLUP is a good thing: The estimation of random effects. *Statistical Science*, **6**: 15-51.

Henderson C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.*, **21**: 309-310.

Henderson C. R. (1963). Selection index and expected genetic advance. *Statistical Genetics and Plant Breeding, National Academy of Science-National Research council Publication*, No. **982**: 141-163.

Henderson C. R. (1984). *Application of Linear Models in Animal Breeding*. U. of Guelph, Canada.

Hill, S. E. and Smith, A. F. M. (1992). Parameterization issues in Bayesian inference. in *Bayesian Statistics 4*. Eds. Bernardo, Berger, David, and Smith. 227-246. Oxford: Oxford University Press.

Jamrozik, J.L., Schaeffer, L. R., and Dekkers J. C. M. (1997). Genetic evaluation of dairy cattle using test day yields and random regression model. *J. Dairy Science*. **80**: 1217-1226

Jensen, J., Mantysaari, E. A., Madsen, P., and Thompson, R. (1996). Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. *J. Ind. Soc. Ag. Statistics*. **97**: 215-236.

Johnson, D. L., and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Science*. **78**: 449-456.

Kackar, R. N., and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.*, **79**: 853-862.

Kirkpatrick, M., and Heckman, N. (1989). A quantitative genetic model for growth, shape, and other infinite-dimensional characters. *J. Math. Biol.*, **27**: 429-450.

Kirkpatrick, M., Hill, W.G., and Thompson, R. (1994). Estimating the covariance structure of traits during growth and aging, illustrated with lactation in dairy cattle. *Genet. Res., Camb.* **64**: 57-69.

Kirkpatrick, M., Lofsvold, D., and Bulmer, M. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*. **124**: 979-993.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*. **38**: 963-974.

Lin, C. Y., and Smith, S. P. (1990). Multi-trait to univariate mixed model analysis of data with multiple random effects *J. Dairy Science*. **73**: 2494-2502.

Lindley, D.V., and Smith, A.F. (1972) Bayesian estimates for the linear model. *J. of Royal Statistical Society*. Series B. **34**: 1-41.

Lindstrom, M.J., and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.*, **83**: 1014-1022.

Littell, R. C., Milliken, G. A., Stroup, W. W. , and Wolfinger, R. D. (1996). *SAS System for Mixed Model.* SAS Inst. Inc., N. Carolina, U.S.A.

Lindstrom, M.J., and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**: 637-687.

McLean, R. A., William, L. S., and Stroup W. W. (1991). A unified approach to mixed linear models. *American Statistician*, **45**: 54-64.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. of Chemical Physics*, **21**: 1087-1092.

Meyer, K. (1989). Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet. Sel. Evol.*, **21**: 317-340.

Meyer, K. (1998). Estimation of covariance functions for longitudinal data using a random regression model. *Genet. Sel. Evol.*, **30**: 221-240.

Meyer, K., and Hill, W. G. (1997). Estimation of genetic and phenotypic covariance function for longitudinal or 'repeated' records by restricted maximum likelihood. *Livestock Prod. Sci.* **47**: 185-200.

Meyer, K., and Smith, S. P. (1996). Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genet. Sel. Evol.*, **28**: 23-49.

Miller, R. J. (1973). Asymptotic properties and computation of Maximum likelihood estimates in the mixed model of the analysis of variance. Technical Report No. 12. Dept. of Statistics, Stanford U., Stanford, California.

Mrode, R. A. (1996). *Linear Models for the Prediction of Animal Breeding Values.* Wallingord: Cab International, U. K.

Norris, J. R. (1997) *Markov Chains.* Cambridge University Press, U. K.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block size are unequal. *Biometrika*, **58**: 545-554.

Ratkowsky, D. A. (1990). *Handbook of Nonlinear Regression Models.* New York: Marcel Dekker.

Ratkowsky, D. A. and Dolby, G. R. (1975). Taylor series linearization and scoring for parameters in nonlinear regression. *J. Applied Statistics*, **24**: 109-111.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (Eds Gilks, Richardson, and Spiegelhalter), 45-58. London: Chapman & Hall.

Robinson, G. K. (1991). That BLUP Is a Good Thing: The estimation of random effects. *Statistical Science*, **6**: 15-51.

Rodriguez-zas, S. L., Gianola, D., and Shook, G. E. (1997). Factors affecting susceptibility to intramamary infection and mastitis: An approximate Bayesian analysis. *J. Dairy Sci.* **80**: 75-85.

Rodriguez-zas, S.L., Gianola, D., and Shook, G.E. (1998). Bayesian analysis of nonlinear mixed effects models for somatic cell score lactation pattern in Holsteins. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production.* **25**: 497-500.

Searle, S. R. (1971). *Linear Models.* New York: John Wiley & Sons.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components.* New York: John Wiley & Sons.

Sheiner, L. B. and Beal, S. L. (1980). Evaluation of methods for estimating population phamacokinetic parameters. I. Michaelis-Menten model: Routine clinical pharmacokinetic data. *J. of Pharmacokinetics and Biopharmaceutics*, **8**: 553-571.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears. *American Stat.*, **46**: 84-88.

Smith, S. P., and Graser, H. U. (1986). Estimating variance components in a class of mixed models by restricted maximum likelihood. *J. Dairy Science*, **69**: 1156-1165.

Strandén, I., and Gianola, D. (1997). Gaussian versus Student-t mixed effects linear models for milk yield in Ayrshire cattle. *European Assoc. Animal Prod. 48th Annual Meeting*, 1-18.

Tanner, M. A. (1993) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood functions*, 2nd edition. New York: Springer-Verlag.

Thisted, R. A. (1988). *Elements of Statistical Computing.* New York: Chapman and Hall.

Tierney L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**: 1701-1762.

Tierney L. (1996). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice* Eds. Gilks, Richardson, and Spiegelhalter, 59-74. London: Chapman & Hall.

Van Vleck, L. D. (1993). *Selection Index and Introduction to Mixed Model Methods.* Florida: CRC Press.

Van Vleck, L. D., and Boldman, K. G. (1993). Sequential transformation for multiple traits for estimation of (co)variance components with a derivative-free algorithm for restricted maximum likelihood. *J. Animal Science*, **71**: 836-844.

van der Werf, J.H.J., Goddard, M. E., and Meyer, K. (1998). The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. *J. Dairy Science*, **81**: 3300-3308.

Varona, L., Moreno, C., Cortes, L.A.G., and Altarriba, J. (1997). Multiple trait genetic analysis of underlying biological variables of production functions. *Livestock Production Sci.*, **47**: 201-209.

Vonesh, E. F., and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, **48**: 1-17.

Vonesh, E. F., and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* New York: Marcel Dekker.

Wakefield J. (1996). The Bayesian analysis of population pharmacokinetic methods. *J. American Stat. Assoc.*, **91**: 62-75.

Wakefield, J. C., Smith, A. F. M., Racine–Poon, A., and Gelfand, A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Appl. Statist.*, **43**: 201-221.

Wright, S. (1922). Coefficients of breeding and relationship. *American Naturalist*, **56**: 330-338.

Wolfinger, R. and Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Comput. Stat. and Data Analysis*, **25**: 465-490.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**: 791-795.

Wright, D. R., Stern, H., and Berger, P. J. (2000). Comparing traditional and Bayesian Analyses of selection experiments in animal breeding. *Agricultural, Biological, and Environmental Statistics*, **5**: 240-256.

©Automatic formatting by APA BIBOUT, a product of Iowa State University Computation Center, designed by Rosanne G. Potter and Joe Struss; original BIBOUT design by Rosanne G. Potter and Steven L. Kunz.